Data science- Recommender system for Amazon

Sadaf Fatollahy

۱۳ دی ۱۴۰۲

نام استاد: دکتر خردپیشه
نام درس : مبانی علوم داده

فهرست مطالب

۱.۰ مقدمه

یک سیستم توصیه توسعه یافته به کسب و کارها کمک می کند تا تجربه خریدار خود را در وب سایت بهبود بخشند و منجر به جذب و حفظ مشتری بهتر شود.Amazon یکی از بزرگترین شرکت های تجارت الکترونیک و رایانش ابری است.آنها ۸.۴ میلیون دلار در آگوست ۲۰۱۳ از دست دادند، زمانی که وب سایت آنها برای ۴۰ دقیقه از کار افتاد. آمازون به شدت به موتور توصیه‌ای متکی است که رتبه‌بندی مشتریان و تاریخچه خرید را برای توصیه اقلام و بهبود فروش بررسی می‌کند.

۲.۰ درباره دیتاست

این مجموعه داده مربوط به بیش از ۲ میلیون بررسی و رتبه بندی مشتری از محصولات مرتبط با زیبایی است که در وب سایت آنها فروخته شده است.

۱. UserId : شناسه کاربری منحصر به فرد مشتری.

۲. ProductId : کد شناسایی منحصر به فرد محصول آمازون برای هر محصول

۳. Ratings : رتبه بندی (از ۱ تا ۵ بر اساس رضایت مشتری)

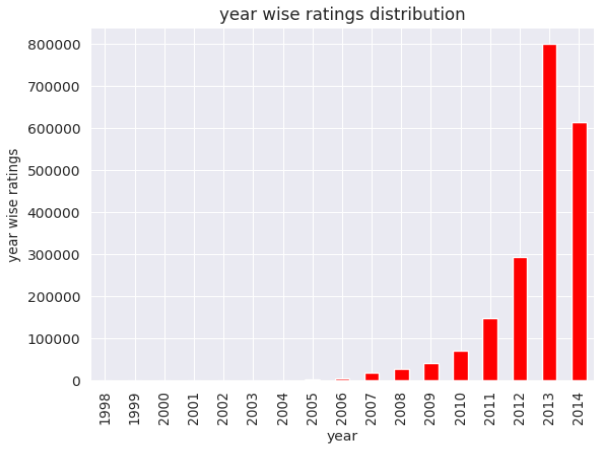
۴. Timestamp : مهر زمانی رتبه بندی (در زمان یونیکس)

۳.۰ EDA and feature engineering

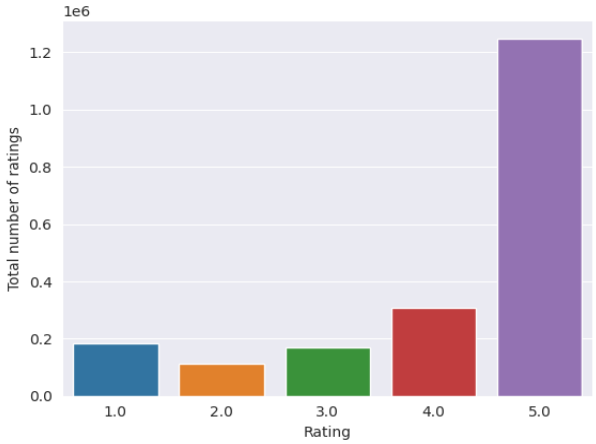
: در ابتدا در دیتاست کمی گشت و گذار کرده و به بررسی فیچر ها میپردازیم.

- این دیتاست شامل ۲۰۲۳۰۷۰ سطر و ۴ ستون است
- میانگین نمرات داده شده به محصولات ۴ بوده است که بیانگر کیفیت بالای محصولاتی است که امازون میفروشد.
- دیتاست missing value ندارد.
- دیتاست داده تکراری ندارد.
- ستون Timestamp بیانگر تعداد ثانیه هایی است که از تاریخ ۱ ژانویه ۱۹۷۰ گذشته است.ما ان را به فرمت قابل فهم تاریخ و ساعت تبدیل میکنیم. تاریخ این دیتا ست از سال ۱۹۹۸ است تا ۲۰۱۴.
- میخواهیم بررسی کنیم میزان نمره دهی به محصولات در طی این سال ها به چه صورت بوده است:

داده‌ها از سال ۱۹۹۸ تا ۲۰۱۴ جمع‌آوری شده‌اند. و ما می‌توانیم هر ساله ببینیم که رتبه‌بندی محصولات ارایشی به طور مداوم در حال افزایش است، به جز افزایش غیرمعمول در سال ۲۰۱۳ به دلایل ناشناخته.

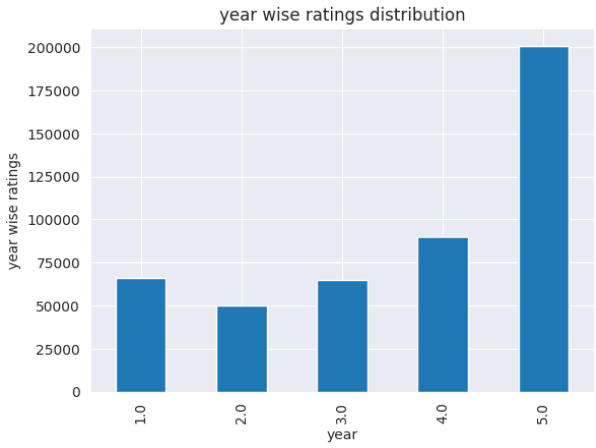


- سپس توزیع rating را بررسی کردیم:



که همانطور که مشخص است بیشتر نمره دهی ۵ بوده که بیانگر کیفیت محصولات عرضه شده میباشد.

- نمودار زیر را درنظر بگیرید:



نمودار بالا نشان میدهد کاربران معمولاً برای تجربه متوسط به محصول نمره نمیدهند، به همین دلیل است که رتبه‌بندی‌های ۲ و ۳ در مقایسه با سایر محصولات، فرکانس کمتری دارند. نمودار شمارش و تعداد محصول منحصربه‌فرد در هر دسته رتبه‌بندی نشان می‌دهد که در رتبه‌بندی ۵ فراوانی بیشتری وجود دارد و پس از آن رتبه‌بندی ۴ و رتبه‌بندی ۱ وجود دارد. این به این معنی است زمانی که کاربران بسیار راضی یا بسیار ناراضی هستند، رتبه‌بندی‌های بیشتری خواهند داد.

۴.۰ Data sparsity

بسیاری از سیستم های توصیه کننده با مشکلی به نام مشکل شروع سرد cold start مواجه می شوند. اساساً به یک کاربر نمی توان چیزی توصیه کرد زیرا به چیزی امتیاز نداده است! علاوه بر این، اگر محصول جدیدی را معرفی شود ، کسی به آن امتیاز نداده و نمی توان آن را توصیه کرد. به این دلیل ، فضای ریاضی نیز بسیار sparse می شود.

به همین دلیل، ما قصد داریم بررسی کنیم که چند رتبه برای هر کاربر ارسال شده است. به این ترتیب میتوان ده کاربری که به محصولات رای داده بودند را پیدا کرد.سپس بررسی کردیم density ماتریس 0.00067 است سپس کاربرانی را که کمتر از ۵۰ بار به محصولات رتبه داده بودند را حذف کردیم به این ترتیب تعداد سطر ها دیتاست به ۲۹۵۵۹ تا کاهش پیدا کرد. دوباره میزان density را بررسی کردیم که برابر با 0.48 شد . ماتریس user-item به صورت ۳۶۱ سطر در ۱۷۲۲۸ ستون تبدیل شد.

۵.۰ Colaborative filtering

User based colaborative filtering و Item based colaborative filtering دو رویکرد متداول هستند که در سیستم های توصیه گر برای ارائه توصیه های شخصی به کاربران استفاده می شوند. آنها از رفتار جمعی و ترجیحات کاربران برای ارائه توصیه ها استفاده می کنند.

- User based colaborative filtering: این روش بر یافتن شباهت‌های بین کاربران بر اساس تعاملات و ترجیحات گذشته آنها تمرکز دارد. فرض بر این است که کاربرانی که سلیقه و ترجیحات مشابهی در گذشته دارند، در آینده نیز ترجیحات مشابهی خواهند داشت.
- Item based colaborative filtering : این روش بر روی یافتن شباهت‌های بین آیتم‌ها بر اساس تعاملات کاربر تمرکز دارد. فرض بر این است که اگر دو مورد به طور مکرر توسط کاربران یکسان رتبه بندی یا با آنها تعامل داشته باشند، احتمالاً از نظر ترجیحات کاربر مشابه هستند.

به این ترتیب ابتدا ماتریس های user-item و item-user را حساب کردیم سپس similarity بین user ها و similarity بین item ها را بررسی کردیم.پس از محاسبه میزان شباهت، زیرمجموعه ای از کاربران مشابه که به ”همسایگی” معروف است، انتخاب می شود. که متشکل از کاربرانی است که بیشترین شباهت را به کاربر هدف دارند. برای ایتم ها نیز به همین صورت عمل کردیم.سپس رتبه‌بندی‌های پیش‌بینی‌شده برای مواردی که کاربر هدف هنوز با آن‌ها تعامل نداشته است، بر اساس رتبه‌بندی کاربران همسایه محاسبه می‌شود. برای ایتم ها نیز به همین صورت است.

نتیجه حاصل برای را برای سه کاربر بر مبنای User based colaborative filtering بررسی کردیم و ۵ ایتم پیشهادی برای هر یک از انها به صورت زیر است:

Recommended Items	user_ratings	user_predictions
B00AE0790U	0.0	2.078306
B00A0J084Y	0.0	1.965524
B00AWLB9I4	0.0	1.887036
B00AQ4EBOI	0.0	1.783238
B008U2Y9BQ	0.0	1.770657

Recommended Items	user_ratings	user_predictions
B000142FVW	0.0	0.783491
B000PQDKU6	0.0	0.675815
B000ZV0D8W	0.0	0.636783
B0069FDR96	0.0	0.618689
B006L1DNWY	0.0	0.609835

Recommended Items	user_ratings	user_predictions
B0056VEYMS	0.0	0.536677
B0092MCO88	0.0	0.526412
B0030HKJ8I	0.0	0.508133
B0000YUX4O	0.0	0.493395
B000ELP5KA	0.0	0.497539