تحلیل داده‌های مربوط به صنعت فیلم‌سازی

نام فریلنس:
فاطمه طاهر

عید ۱۴۰۳

چکیده

پیش‌بینی سودآوری فیلم با هدف توسعه یک مدل برای تعیین اینکه آیا یک فیلم بر اساس ویژگی‌های مختلف مانند بودجه، درآمد، ژانر، شرکت تولید و غیره سودآور خواهد بود یا خیر. پروژه شامل مراحلی از جمله اکتشاف داده، پیش‌پردازش، مهندسی ویژگی، مدل‌سازی و ارزیابی می‌باشد. نکات کلیدی عبارتند از:
کاوش جامع داده‌ها، بینش‌هایی را در مورد توزیع ویژگی‌های کلیدی و الگوهای شناسایی شده نشان داد.
پیش‌پردازش داده شامل مدیریت مقادیر ازدست‌رفته، پاکسازی داده‌ها و تغییر ویژگی‌های دسته ای است.
مهندسی ویژگی، ویژگی‌های جدیدی ایجاد کرد و ویژگی‌های موجود را برای بهبود قدرت پیش‌بینی اصلاح کرد.
مدل‌های مختلف یادگیری ماشین از جمله رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی و XGBoost آموزش و ارزیابی شدند.
جنگل تصادفی به عنوان بهترین مدل با بالاترین درصد دقت ظاهر شد.
این پروژه اهمیت کاوش کامل داده‌ها، پیش‌پردازش، و مهندسی ویژگی‌ها را در ساخت مدل‌های پیش‌بینی برای سودآوری فیلم نشان داد. کارهای آینده شامل اصلاح بیشتر تکنیک‌های مهندسی ویژگی و آزمایش با تکنیک‌های مدل‌سازی پیشرفته برای بهبود عملکرد پیش‌بینی می‌شود.

واژه‌های کلیدی: تحلیل داده، رگرسیون، شناسایی الگو، سیستم‌هایی توصیه‌دهنده، فیلم.

فهرست مطالب

فهرست اشکال

فهرست جداول

مقدمه

در عصری که موفقیت یک فیلم می‌تواند به طور قابل توجهی بر صنعت سرگرمی تأثیر بگذارد، درک عواملی که در سودآوری یک فیلم نقش دارند بسیار مهم است. پروژه پیش‌بینی سودآوری فیلم به این حوزه می‌پردازد و هدف آن پیش‌بینی موفقیت مالی فیلم قبل از اکران آن است. با تجزیه و تحلیل طیف متنوعی از ویژگی‌ها از جمله بودجه، سال ساخت، ژانر، شرکت تولید و موارد دیگر، این پروژه به دنبال کشف الگوها و روندهایی است که بر سودآوری بالقوه فیلم تأثیر می‌گذارد. از طریق کاوش جامع داده‌ها، پیش‌پردازش، مهندسی ویژگی و مدل‌سازی، این پروژه تلاش می‌کند تا بینش‌های ارزشمندی را برای ذینفعان صنعت فیلم فراهم کند و به فرآیندهای تصمیم‌گیری و برنامه‌ریزی استراتژیک کمک کند.

تحلیل و بررسی مجموعه داده:

مجموعه داده مورد استفاده در پروژه پیش‌بینی سودآوری فیلم، مجموعه‌ای جامع از اطلاعات پنج هزار فیلم را ارائه می‌دهد.

۱. ویژگی‌ها: مجموعه داده شامل طیف گسترده‌ای از ویژگی‌ها از جمله:

- بودجه: مبلغی که برای ساخت فیلم اختصاص داده شده است.
- درآمد: درآمد حاصل از فیلم پس از اکران آن.
- ژانر: دسته یا نوع فیلم (به عنوان مثال، اکشن، کمدی، درام).
- شرکت سازنده: شرکت سازنده فیلم.
- تاریخ انتشار: تاریخ اکران فیلم.
- محبوبیت: معیاری که میزان محبوبیت فیلم را نشان می‌دهد.
- زمان اجرا: مدت زمان فیلم بر حسب دقیقه.
- میانگین رای: میانگین امتیازی که بینندگان به فیلم می‌دهند.
- تعداد آرا: تعداد آرا/امتیازهای دریافت شده توسط فیلم.
- زبان: زبان اصلی فیلم.
- کشور: کشوری که فیلم در آن تولید یا فیلمبرداری شده است.

- وضعیت: وضعیت فعلی فیلم (به عنوان مثال، منتشر شده، در حال تولید).

۲. کیفیت داده: به نظر می‌رسد مجموعه داده نسبتاً تمیز با حداقل مقادیر گمشده و ناسازگاری است. با این حال، قبل از انجام هر گونه تجزیه و تحلیل یا مدل‌سازی، تمیز کردن و پیش‌پردازش کامل داده‌ها ضروری است.

۳. توزیع داده‌ها: تجزیه و تحلیل اولیه ویژگی‌های عددی مانند بودجه، درآمد و زمان اجرا، طیف وسیعی از مقادیر را با مقادیر پرت بالقوه نشان می‌دهد. برای درک توزیع این ویژگی‌ها و شناسایی هر گونه ناهنجاری به کاوش بیشتر نیاز است.

۴. ویژگی‌های طبقه‌بندی: ویژگی‌های دسته‌بندی مانند ژانر، شرکت سازنده، زبان و کشور، بینش‌های ارزشمندی را در مورد تنوع فیلم‌ها در مجموعه داده ارائه می‌دهد. تجزیه و تحلیل توزیع فراوانی این دسته‌ها می‌تواند به کشف روندها و الگوها کمک کند.

۵. روند زمانی: ویژگی تاریخ انتشار امکان تجزیه و تحلیل روندهای زمانی در صنعت فیلم را فراهم می‌کند. درک اینکه چگونه عواملی مانند سال اکران و فصلی بودن بر سودآوری فیلم تأثیر می‌گذارد، می‌تواند بینش ارزشمندی برای تصمیم‌گیری ارائه دهد.

۶. محبوبیت و رتبه‌بندی: ویژگی‌هایی مانند محبوبیت، میانگین آرا و تعداد آرا نشان‌دهنده درگیری و استقبال مخاطبان از فیلم‌ها است. تجزیه و تحلیل این معیارها می‌تولند ترجیحات بینندگان را آشکار کند و به پیش‌بینی موفقیت بالقوه فیلم کمک کند.

۷. چالش‌های بالقوه: ممکن است در رسیدگی به متغیرهای طبقه بندی شده با کاردینالیتی بالا (مثلاً شرکت های تولید) و برخورد با داده‌های گم‌شده یا ناقص، چالش‌هایی ایجاد شود. علاوه بر این، موارد پرت و ناسازگاری در ویژگی‌های عددی ممکن است نیاز به تصحیح دقیق در طول پیش‌پردازش داده‌ها داشته باشد.

به طور کلی، مجموعه داده منبعی غنی از اطلاعات برای بررسی عوامل موثر بر سودآوری فیلم ارائه می‌دهد.

تحلیل اکتشافی داده‌ها:

این مجموعه داده دارای تنها *سه مقدار گمشده* می‌باشد که آن‌ها را حذف می‌کنیم.
رمزگذاری شمارشی، تکنیکی است که برای نمایش متغیرهای طبقه‌بندی‌شده با تعداد فراوانی آنها که در مجموعه داده استفاده شده است. در اینجا توضیح مختصری درباره رمزگذاری تعداد برای *ژانرها، شرکت‌های تولید، کشورهای سازنده و زبان‌های گفتاری* آمده است:

۱. ژانرها: رمزگذاری تعداد ژانرها شامل جایگزینی هر دسته ژانر با تعداد فیلم‌های متعلق به آن ژانر است. به عنوان مثال، اگر«اکشن» ۱۰۰ بار در مجموعه داده ظاهر شود، تمام رخدادهای «اکشن» با مقدار ۱۰۰ جایگزین می‌شوند.

۲. شـرکت‌های تولیدکننده: به طور مشـابه، رمزگذاری تعداد برای شـرکت‌های تولید شـامل جایگزینی هر شـرکت تولیدکننده با تعداد فیلم‌های مرتبط با آن شـرکت اسـت. این رمزگذاری نشـان‌دهنده فراوانی شرکت‌های تولیدی در مجموعه داده است.

۳. کشـورهای تولید: کدگذاری شـمارش برای کشـورهای سـازنده از یک اصـل پیروی می‌کند، جایی که تعداد فیلم‌های تولید شـده در آن کشـور جایگزین هر کشـور می‌شـود. این رمزگذاری توزیع مکان‌های تولید را در سراسر مجموعه داده ثبت می‌کند.

۴. زبان‌های گفتاری: رمزگذاری تعداد برای زبان‌های گفتاری مسـتلزم جایگزینی هر زبان با تعداد فیلم‌های صحبت شده به آن زبان است. این نشان‌دهنده شیوع هر زبان در مجموعه داده است.

رمزگذاری شـمارش روشـی سـاده و در عین حال موثر برای گنجاندن اطلاعات طبقه‌بندی شـده در مدل‌های یادگیری ماشـین اسـت. نظم دسـته‌ها را حفظ می‌کند در حالی که فرکانس‌های نسـبی آن‌ها را در مجموعه داده ثبت می‌کند، و آن را به یک تکنیک پیش‌پردازش مفید برای وظایف مدل‌سـازی پیش‌بینی‌کننده تبدیل می‌کند.

نقشـه برداری[1] از *زبان اصـلی* شـامل تبدیل زبان اصـلی فیلم‌ها به نمایش‌های عددی اسـت. به هر زبان منحصربه‌فرد یک کد عددی خاص اختصاص داده می‌شود که به الگوریتم‌های یادگیری ماشین اجازه می‌دهد تا ویژگی زبان را به طور موثر پردازش و تجزیه و تحلیل کنند.

نرمال‌سازی[2] *بودجه، درآمد و شمارش رأی* شامل مقیاس‌بندی این ویژگی‌های عددی در محدوده بین ۰ و ۱ اسـت. این معمولاً با کم کردن مقدار حداقل از هر نقطه داده و سـپس تقسـیم بر دامنه (یعنی تفاوت بین حداکثر و حداقل) انجام می‌شـود. نرمال‌سـازی این ویژگی‌ها تضـمین می‌کند که آنها در یک مقیاس ثابت هسـتند، که می‌تواند عملکرد الگوریتم های یادگیری ماشـین را که به بزرگی متغیرهای ورودی حسـاس هستند، بهبود بخشد.
تفسیر همبستگی بین ویژگی‌ها در مجموعه داده برای درک روابط و شناسایی الگوها یا وابستگی‌های بالقوه بسیار مهم است. نحوه تفسیر نتایج حاصل از تجزیه و تحلیل همبستگی شامل:

۱) همبسـتگی مثبت: اگر ضـریب همبسـتگی نزدیک به ۱+ باشـد، نشـان‌دهنده رابطه خطی مثبت قوی بین دو ویژگی اسـت. این بدان معناسـت که با افزایش یک ویژگی، ویژگی دیگر نیز تمایل به افزایش دارد. برای مثال، همبسـتگی مثبت بین «بودجه» و «درآمد» نشـان می‌دهد که فیلم‌ها با بودجه بالاتر معمولا درآمد بیشتری دارند.

۲) همبستگی منفی: اگر ضریب همبستگی نزدیک به ۱- باشد، نشان‌دهنده رابطه خطی منفی قوی بین دو ویژگی اسـت. این بدان معنی اسـت که با افزایش یک ویژگی، ویژگی دیگر کاهش می‌یابد. برای مثال، بین «سـال انتشـار» و «زمان اجرا» همبسـتگی منفی وجود دارد، که نشـان‌می‌دهد که فیلم‌های قدیمی‌تر ممکن

[1] Mapping
[2] Normalization

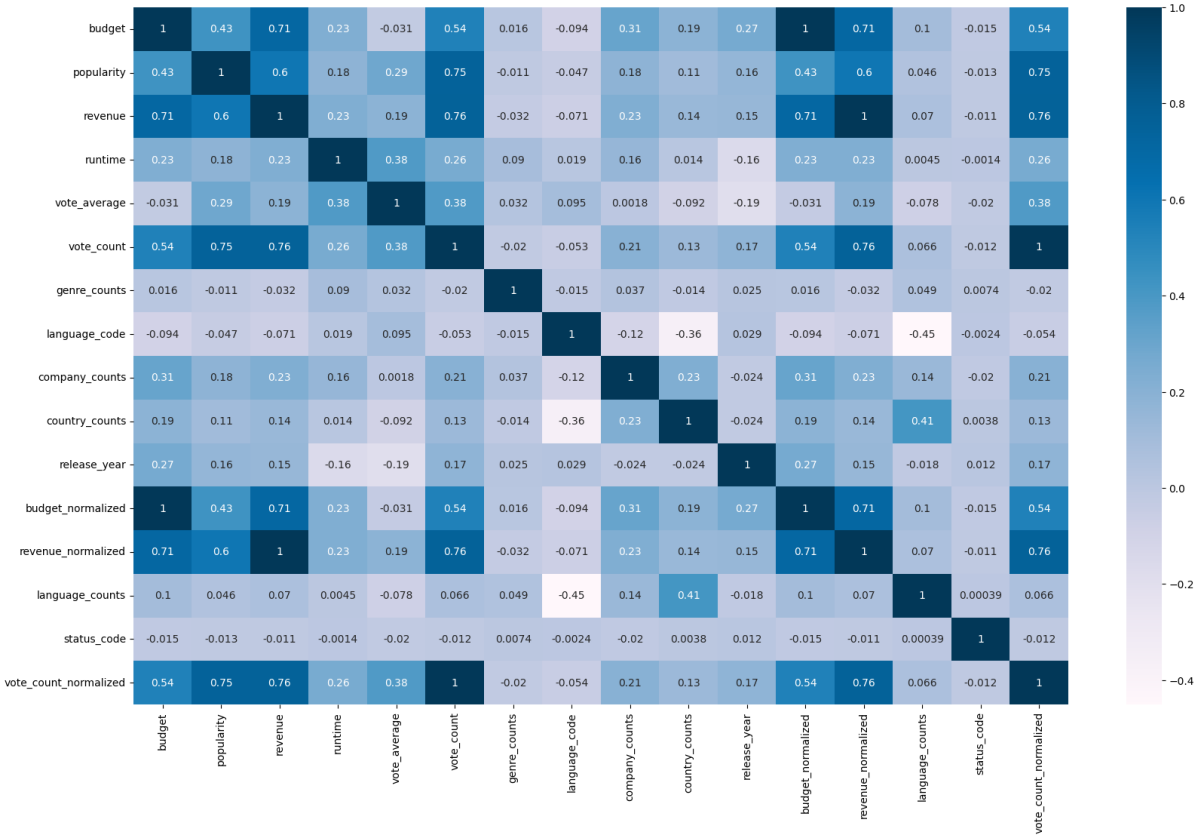
است در مقایسه با فیلم‌های جدیدتر زمان اجرای کمتری داشته باشند.

۳) همبستگی ضعیف: اگر ضریب همبستگی نزدیک به ۰ باشد، نشان دهنده وجود رابطه خطی ضعیف یا عدم وجود رابطه خطی بین دو ویژگی است. در این حالت، تغییرات در یک ویژگی لزوما منجر به تغییر در ویژگی دیگر نمی‌شود. با این حال، توجه به این نکته مهم است که هنوز هم می‌تواند روابط غیرخطی یا غیر یکنواختی وجود داشته باشد که توسط ضریب همبستگی ثبت نشده باشد.

۴) بدون همبستگی: ضریب همبستگی ۰ نشان‌دهنده عدم وجود رابطه خطی بین دو ویژگی است. این لزوما به این معنا نیست که اصلا رابطه‌ای وجود ندارد. به سادگی به این معنی است که هیچ ارتباط خطی بین ویژگی‌ها وجود ندارد.

۵) تفسیر: بر اساس تحلیل همبستگی، می‌توانیم در مورد تاثیرگذاری و سمت روابط بین جفت ویژگی‌ها نتیجه‌گیری کنیم. این نتیجه‌گیری‌ها می‌توانند به تحلیل بیشتر یا فرآیندهای تصمیم‌گیری کمک کنند. برای مثال، اگر یک همبستگی مثبت قوی بین «بودجه» و «درآمد» پیدا کردیم، ممکن است سرمایه‌گذاری در فیلم‌های با بودجه بالاتر را برای به حداکثر رساندن درآمد بالقوه در اولویت قرار دهیم.

۶) علیت: توجه به این نکته مهم است که همبستگی دلالت بر علیت ندارد. حتی اگر دو ویژگی به شدت با هم مرتبط باشند، لزوما به این معنی نیست که تغییرات در یک ویژگی باعث تغییر در ویژگی دیگر می‌شود. ممکن است برای ایجاد روابط علّی بین ویژگی‌ها، تحلیل‌های اضافی، مانند آزمایش‌ها یا تکنیک‌های استنتاج علّی، مورد نیاز باشد.



شکل ۱- ضریب همبستگی بین متغیرهای مجموعه داده

آموزش و ارزیابی مدل‌ها:

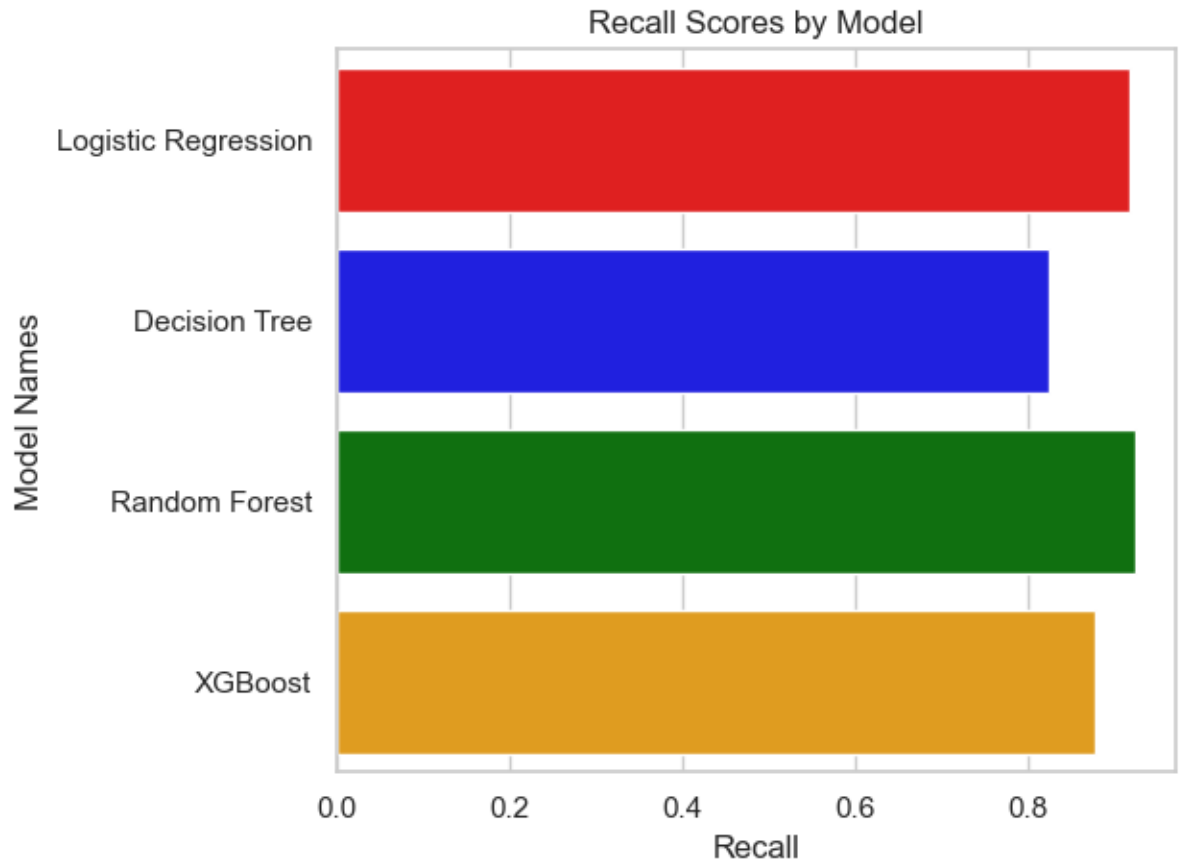
برای ایجاد مدل ما به برچسبی مناسب برای سنجش میزان سودآوری یک فیلم احتیاج داریم که این برچسب یا ویژگی را با مقایسه میزان بودجه و درآمد هر فیلم بدست می‌آوریم. در جایی که میزان بودجه از درآمد بیشتر یا برار بوده یعنی فیلم موردنظر، سودآور نمی‌باشد پس برچسب ۰ را برای آن در نظر می‌گیریم؛ در حالت عکس این قضیه برچسب ما ۱ خواهد بود. پس دو کلاس برای طبقه‌بندی وجود دارد.

برای تقسیم‌بندی داده‌ها ۸۰ درصد به عنوان داده‌های آموزش و ۲۰ درصد داده‌ها برای آزمایش اختصاص داده شده‌اند و برای مقیاس‌بندی و نرمال‌سازی ویژگی‌ها، داده‌ها به میانگین صفر و واریانس واحد نرمال شده‌اند.

با تجزیه و تحلیل این داده‌ها، می‌توان الگوهایی را شناسایی کرد که به ما کمک می‌کند تا بفهمیم چه ویژگی‌هایی، سودآوری یک فیلم را بیشتر می‌کنند. همچنین می‌توان این داده‌ها را برای ساخت مدل‌های پیش‌بینی استفاده کرد که می‌توانند با دقت قابل قبولی پیش‌بینی کنند میزان سودآوری چقدر خواهد بود. پس از پاکسازی و نرمال‌سازی داده‌ها اقدام به آموزش مدل‌های یادگیری ماشین و بدست اوردن معیارهای ارزیابی مدل‌ها می‌کنیم. الگوریتم‌های مورد استفاده جهت شناسایی الگوها عبارت‌اند از الگوریتم‌های مختلفی همچون رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، XGBoost که معیارهای ارزیابی دقت[1]، صحت[2]، یادآوری[3]، امتیاز F1[4] و ناحیه زیر منحنی[5] برای هر کدام از آنها محاسبه شده است.

---

[1] Accuracy
[2] Precision
[3] Recall
[4] F1-Score
[5] ROC



شکل ۲- مقدار Recall برای الگوریتم‌های مختلف

همانطور که در شکل بالا مشاهده می‌شود مقدار Recall برای الگوریتم لجستیک خطی از دیگر الگوریتم‌ها بیشتر است که این امر نشان‌دهنده عملکرد بهتر و قدرت پیش‌بینی بالاتر این مدل می باشد. مقادیر میانگین مربعات خطا، جذر میانگین مربعات خطا و ضـــریب تعیین برای الگوریتم های مختلف در جدول زیر نمایش داده شده اند.

جدول۱ - معیارهای ارزیابی برای الگوریتم‌های مختلف

Algorithms	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.756966	0.793286	0.918200	0.851185	0.586489
Decision Tree	0.738390	0.829218	0.824131	0.826667	0.647734
Random Forest	0.797214	0.827839	0.924335	0.873430	0.662805
XGBoost	0.777090	0.836257	0.877301	0.856287	0.671134

سیستم‌های توصیه:

در اینجا یک نمای کلی از سه طرح سیستم توصیه محبوب آورده شده است:

۱) فیلتر جمعیت شناختی (توصیه‌کننده ساده):
سیستم فیلم‌های مشابه را به کاربرانی با ویژگی‌های جمعیتی مشابه توصیه می‌کند. این سیستم‌ها بر اساس محبوبیت فیلم و/یا ژانر، توصیه‌های کلی را به هر کاربر ارائه می‌دهند. از آنجایی که هر کاربر متفاوت است، این رویکرد بسیار ساده در نظر گرفته می‌شود. ایده اصلی پشت این سیستم این است که فیلم‌هایی که محبوبیت بیشتری دارند و مورد تحسین منتقدان قرار می‌گیرند، احتمال بیشتری برای دوست‌داشتن بینندگان معمولی دارند.

۲) فیلتر مشارکتی:
فیلتر مشارکتی موارد را با استفاده از اولویت‌های سایر کاربران توصیه می‌کند. فرض بر این است که کاربرانی که در گذشته توافق کرده‌اند، در آینده مجدداً موافقت خواهند کرد.
دو نوع اصلی فیلتر مشارکتی عبارتند از:
فیلتر مشارکتی مبتنی بر کاربر: یافتن کاربران مشابه با کاربر هدف و توصیه مواردی که آنها دوست داشته‌اند، به کاربر هدف.
فیلتر مشارکتی مبتنی بر آیتم: مواردی، مشابه مواردی که کاربر هدف در گذشته دوست داشته است، را توصیه می‌کند.
فیلتر مشارکتی نیازی به استخراج ویژگی یا درک محتوای مورد ندارد. در عوض، تنها بر تعاملات کاربر-آیتم تمرکز دارد.
چالش‌ها شامل مشکل شروع سرد (برای کاربران/موارد جدید) و مقیاس‌پذیری تنها با مجموعه داده‌های بزرگ میسر است.

3) فیلترینگ مبتنی بر محتوا:
فیلتر مبتنی بر محتوا مواردی مشابه مواردی را که کاربر در گذشته دوست داشته است، بر اساس محتوا/ویژگی‌های موارد توصیه می‌کند.
این امر متکی بر استخراج ویژگی‌ها از آیتم‌ها (مانند ژانرهای فیلم، بازیگران، کارگردانان) و یافتن شباهت‌ها بین آیتم‌ها بر اساس این ویژگی‌ها است. فیلتر مبتنی بر محتوا می‌تواند مشکل شروع سرد را بهتر از فیلتر مشترک حل کند، زیرا به تعاملات کاربر متکی نیست.
با این حال، به ابرداده‌های غنی مربوط به آیتم نیاز دارد و ممکن است دچار تخصص بیش‌ازحد شود، در جایی که به کاربران بارها موارد مشابه توصیه می‌شود.

در این پروژه، سیستم توصیه مورد استفاده یک رویکرد فیلترینگ مبتنی بر محتوای ترکیبی است.

۱. استخراج ویژگی: ویژگی‌های مرتبط مانند ژانرهای فیلم، بازیگران و گروه تولید از مجموعه داده استخراج

می‌شوند.

۲. بردارسازی ویژگی: هر فیلم به عنوان یک بردار ویژگی بر اساس این ویژگی‌های استخراج شده نشان داده می‌شود. این یک نمایش عددی از فیلم‌ها را ایجاد می‌کند که شامل اطلاعاتی درباره ژانرها، بازیگران و گروه تولید است.

۳. محاسبه شباهت: شباهت بین فیلم‌ها با استفاده از معیار تشابه مانند شباهت کسینوس محاسبه می‌شود. این میزان شباهت یک فیلم به فیلم دیگر را بر اساس بردارهای ویژگی آن، که شامل اطلاعاتی درباره ژانرها، بازیگران و گروه تولید است، می‌سنجد.

۴. توصیه‌گر: برای یک فیلم معین، سیستم مشابه‌ترین فیلم‌ها را بر اساس شباهت‌های محاسبه شده پیدا می کند. سپس این فیلم‌های مشابه بر اساس امتیاز شباهتشان مرتب می‌شوند و ۵ فیلم‌های برتر به کاربر توصیه می‌شوند.

به طور کلی، سیستم توصیه فیلتر مبتنی بر محتوای ترکیبی از اطلاعات مربوط به ژانرهای فیلم، بازیگران و گروه تولید برای تولید توصیه‌ها استفاده می‌کند و برای مواردی که ترجیحات کاربر مشخص است و ویژگی‌های آیتم‌ها در دسترس است مناسب است.