به نام خدا

درس علم داده

گزارش پروژه: پیشبینی درآمد فیلمها و تحلیل اکتشافی داده ها

اعضای گروه: محمدرضا ادریس آبادی (۹۹۵۲۲۳۶۵) ، پارسا آقاعلی(۴۰۰۵۲۱۰۷۲)

استاد: دکتر نادری

پاییز ۱۴۰۳

مقدمه

در دنیای سینما، پیش‌بینی میزان فروش یک فیلم در گیشه یکی از چالش‌های اساسی برای استودیوهای فیلم‌سازی و سرمایه‌گذاران محسوب می‌شود. عوامل مختلفی همچون بودجه تولید، ژانر فیلم، تعداد بازیگران، سابقه کارگردان، امتیاز منتقدان و بینندگان، و حتی زمان اکران می‌توانند بر درآمد یک فیلم تأثیرگذار باشند. هدف این پروژه، بررسی و تحلیل این متغیرها و طراحی مدل‌های یادگیری ماشین برای پیش‌بینی درآمد فیلم‌ها بر اساس داده‌های **Rotten Tomatoes** است.

این گزارش شامل مراحل مختلف پردازش داده، مهندسی ویژگی‌ها، آموزش و مقایسه مدل‌های یادگیری ماشین، تحلیل داده‌ها و پیشنهادهایی برای بهبود مدل خواهد بود.

۲ .آماده‌سازی و پیش‌پردازش داده‌ها

۲.۱ بارگذاری داده‌ها

برای انجام این تحلیل، دو مجموعه داده شامل اطلاعات فیلم‌ها و اطلاعات عوامل (بازیگران و کارگردانان) مورد استفاده قرار گرفتند. این داده‌ها از دو فایل **CSV**بارگذاری و با استفاده از pandasپردازش شدند.

۲.۲ پردازش داده‌های متنی و دسته‌بندی

- **تبدیل مقادیر متنی به عددی :**برخی متغیرهای کیفی (مانند زبان اصلی فیلم) با استفاده از LabelEncoderبه مقادیر عددی تبدیل شدند.
- **ژانر فیلم‌ها (One-Hot Encoding) :**ژانر فیلم‌ها که به‌صورت لیستی از مقادیر متنی ذخیره شده بودند، با روش **One-Hot Encoding**به متغیرهای عددی مجزا تبدیل شدند.
- **تبدیل تاریخ اکران :** تاریخ انتشار فیلم‌ها به سال انتشار تبدیل شد تا بتوان از آن به‌عنوان یک ویژگی عددی استفاده کرد.

۲.۳ مدیریت داده‌های گمشده

- **حذف سطرهای دارای مقادیر گمشده مهم :**مواردی که مقدار گمشده آن‌ها اهمیت بالایی داشت (مانند rt_box_office)حذف شدند.

- **جایگذاری مقادیر گمشده با مقادیر میانه :**برای ویژگی‌هایی مانند rt_runtime،

rt_production_budgetو rt_review_count، مقادیر گمشده با میانه پر شدند تا توزیع

داده‌ها حفظ شود.

۳ .مهندسی ویژگی‌ها

۳.۱ ایجاد ویژگی‌های جدید

- **تعداد بازیگران :**تعداد بازیگرانی که در یک فیلم حضور دارند به‌عنوان یک ویژگی اضافه شد.
- **شمارش ژانرها :**تعداد ژانرهایی که یک فیلم به آن تعلق دارد محاسبه شد.
- **کارگردانان شناخته‌شده :**بررسی شد که آیا کارگردان فیلم از بین افراد شناخته‌شده است یا خیر.
- **کشورهای تولیدکننده :**کشورهایی که در تولید فیلم مشارکت داشته‌اند به‌عنوان یک ویژگی دسته‌ای

بررسی و پردازش شدند.

۴ .آموزش مدل‌ها

۴.۱ تقسیم داده‌ها به مجموعه آموزشی و تست

داده‌ها به دو مجموعه آموزشی (٪۷۵) و تست (٪۲۵) تقسیم شدند. ویژگی‌های عددی نرمال‌سازی شدند و مدل‌ها بر روی این داده‌ها آموزش دیدند.

۴.۲ مدل‌های مورد استفاده

- **رگرسیون خطی (Linear Regression)** مدلی ساده که رابطه خطی بین ویژگی‌ها و درآمد را تحلیل می‌کند.
- **جنگل تصادفی Random Forest Regressor)** مدلی غیربرخطی که با استفاده از مجموعه‌ای از درخت‌های تصمیم، دقت پیش‌بینی را افزایش می‌دهد.
- **XGBoost** مدلی مبتنی بر درخت‌های افزایشی که دقت بالاتری نسبت به روش‌های کلاسیک دارد.

۴.۳ تنظیم هایپرپارامترها

برای بهینه‌سازی مدل‌ها، تنظیمات زیر برای **XGBoost**اعمال شد:

- n_estimators: 200
- learning_rate: 0.1
- max_depth: 6

مدل‌های دیگر نیز با مقداردهی پیش‌فرض اجرا شدند.

۵ .ارزیابی مدل‌ها

۵.۱ معیارهای ارزیابی

مدل‌ها با استفاده از معیارهای زیر ارزیابی شدند:

- **میانگین قدر مطلق خطا (MAE)** مقدار مطلق خطای پیش‌بینی‌های مدل.
- **میانگین مربعات خطا (MSE)** مربعات خطای پیش‌بینی‌ها.
- **ضریب تعیین (R^2)** معیاری برای سنجش دقت مدل.

۵.۲ مقایسه عملکرد مدل‌ها

نتایج ارزیابی مدل‌ها به‌صورت زیر بود:

مدل	MAE	MSE	R^2
جنگل تصادفی	۳۸,۸۷۶,۳۴۵	$۱۰^{15} \times ۷.۹۹۲$	۰.۷۵۳
XGBoost	۴۰,۰۹۹,۸۸۹	$۱۰^{15} \times ۹.۰۵۳$	۰.۷۲۱

بر اساس این مقایسه، **جنگل تصادفی** بهترین عملکرد را در میان مدل‌های بررسی‌شده داشت.

۶ .تحلیل داده‌ها

۶.۱ نمودار مقایسه مقادیر واقعی و پیش‌بینی‌شده

نموداری برای مقایسه درآمد واقعی فیلم‌ها با پیش‌بینی‌های مدل XGBoost رسم شد. نقاطی که به خط ایده‌آل $y=x$ نزدیک‌تر باشند، نشان‌دهنده دقت بالاتر مدل هستند.

۶.۲ بررسی اهمیت ویژگی‌ها

ویژگی‌های مهم که تأثیر بیشتری در پیش‌بینی درآمد فیلم داشتند عبارت بودند از:

۱. **تعداد نقدهای ثبت‌شده(rt_review_count)**

۲. **بودجه تولید(rt_production_budget)**

۳. **کشور تولیدکننده(rt_production_countries)**

۴. **امتیاز منتقدان و تماشاگران rt_critics_score) و(rt_audience_score**

۵. **تعداد بازیگران(rt_actors_count)**

نمودار اهمیت ویژگی‌ها نشان داد که تعداد نقدها و بودجه تولید تأثیرگذارترین عوامل در پیش‌بینی درآمد فیلم هستند.

۷ .پیشنهادات برای بهبود مدل

- افزایش حجم داده‌ها :داده‌های بیشتری از سایر منابعIMDb ، BoxOfficeMojo می‌توانند به مدل اضافه شوند.
- استفاده از مدل‌های پیچیده‌تر :مانند Neural Networks برای افزایش دقت.
- بهینه‌سازی هایپرپارامترها : به‌ویژه برای مدل Random Forest، استفاده از GridSearchCVیا Bayesian Optimization.

- اضافه کردن ویژگی‌های جدید :مانند بازیگران اصلی فیلم، نقدهای کاربران، و تأثیر تبلیغات.

۸ .نتیجه‌گیری

در این پروژه، مدل‌های یادگیری ماشین برای پیش‌بینی درآمد فیلم‌های سینمایی مورد بررسی قرار گرفتند. نتایج نشان داد که مدل **جنگل تصادفی** بهترین عملکرد را داشته و معیار R^2 آن برابر **۰.۷۵** بود. همچنین، مشخص شد که **تعداد نقدها، بودجه تولید و کشور تولیدکننده** بیشترین تأثیر را در پیش‌بینی درآمد فیلم دارند.

در آینده، **بهبود مدل با استفاده از داده‌های بیشتر و الگوریتم‌های پیشرفته‌تر** می‌تواند دقت پیش‌بینی را افزایش دهد و ابزارهای دقیق‌تری برای تحلیل و پیش‌بینی فروش فیلم‌ها ارائه دهد.

منابع

- **Scikit-learn Documentation**: https://scikit-learn.org/stable/
- **XGBoost Documentation**: https://xgboost.readthedocs.io/en/latest/
- **Rotten Tomatoes Dataset**: https://www.rottentomatoes.com/
- **IMDb Dataset**: https://www.imdb.com/interfaces/
- **Box Office Mojo**: https://www.boxofficemojo.com/