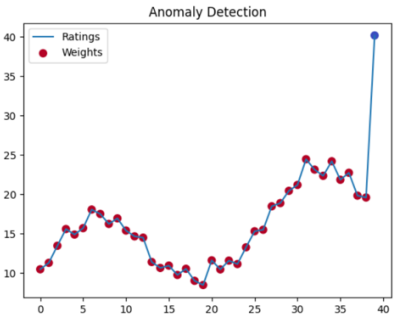
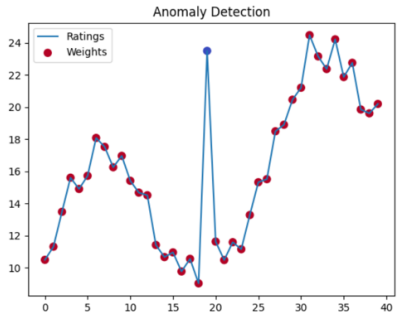
مستندات مدیریت داده‌های نامتعارف

در این پروژه، به منظور بهینه‌سازی مدیریت داده‌ها، تمامی امتیازات ثبت‌شده توسط کاربران (جدول UserRating) به صورت روزانه یک سطر در جدول RatingBin را میسازند. هر سطر در این جدول شامل میانگین امتیازات و تعداد کاربرانی است که در آن روز به یک مطلب خاص امتیاز داده‌اند. امتیاز نهایی هر مطلب (جدول Blog) بر اساس اطلاعات موجود در این جدول محاسبه و به‌روزرسانی می‌شود.

استراتژی محاسبه امتیازات:
برای محاسبه امتیاز نهایی هر مطلب، به هر سطر در جدول RatingBin یک وزن تخصیص داده می‌شود. اگر تعداد امتیازدهندگان یک روز به طور غیرمعمولی نسبت به روزهای دیگر بیشتر باشد، به آن سطر وزن کمتری داده می‌شود.

تشخیص داده‌های نامتعارف:
برای شناسایی روزهایی که تعداد امتیازدهندگان یک مطلب نسبت به روزهای دیگر غیرعادی است، از کتابخانه‌ی statsmodels استفاده شده است. این کتابخانه امکان شناسایی داده‌های پرت را فراهم می‌کند. در این روش، تعداد امتیازدهندگان هر سطر از جدول RatingBin به عنوان ورودی داده می‌شود. این ابزار به هر مقدار یک مقدار عددی تخصیص می‌دهد که نشان‌دهنده میزان همخوانی آن با روند کلی است. اگر این مقدار به طور قابل‌توجهی (به طور پیش‌فرض بیشتر از ۳ برابر انحراف معیار) بیشتر باشد، به آن وزن ۰.۱ (به طور پیش‌فرض) اختصاص داده می‌شود.





در اینجا نقاط آبی نشان دهنده ی نقاط غیر متعارف هستند، که در محاسبه میانگین هر مطلب وزن کمتری میگیرند.

نتیجه نهایی:
این استراتژی به بهبود دقت در محاسبه امتیازات نهایی کمک کرده و از تأثیرگذاری داده‌های نامتعارف یا روزهای با رفتار غیرعادی بر امتیاز نهایی جلوگیری می‌کند.