Data science- Book price prediction

Sadaf Fatollahy

۱۰ آذر ۱۴۰۲

نام استاد: دکتر خردپیشه
نام درس : مبانی علوم داده

فهرست مطالب

۱.۰ مقدمه

در این تمرین هدف ما پیش بینی قیمت های کتاب بر اساس فیچر های مشخصی است که در ادامه ذکر میکنیم. در این تمرین ۲ دیتاست جداگانه برای train و test داریم که به طور موازی روی هردو preprocessing انجام میدهیم. برای این تمرین ما از مدل random forest regressor و همجنین از معیار MSE استفاده کردیم.

۲.۰ درباره دیتاست

دیتاست train شامل ۵۶۹۹ سطر و ۹ ستون است و دیتاست test شامل۵۳۷ سطر و ۹ ستون است .حال بررسی میکنیم هر فیچر نشان دهنده چیست؟

۱. Title :عنوان کتاب

۲. Author :نویسنده (نویسندگان) کتاب.

۳. Edition:نسخه کتاب

۴. Reviews :نظرات مشتری در مورد کتاب

۵. Ratings :رتبه‌بندی مشتریان کتاب

۶. Synopsis: خلاصه کتاب

۷. Genre :ژانری که کتاب به آن تعلق دارد

۸. BookCategory :بخشی که کتاب معمولاً در آن موجود است.

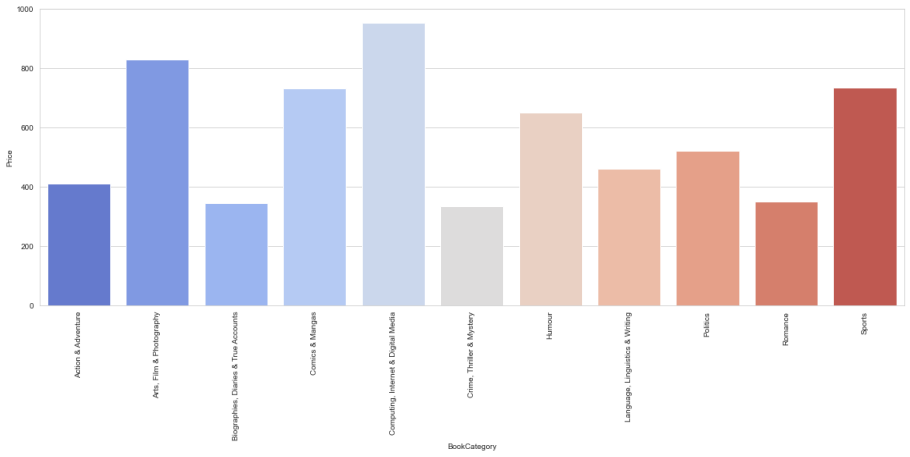
۹. Price: قیمت کتاب (متغیر هدف)

۳.۰ EDA , Data Cleaning and Feature Engineering

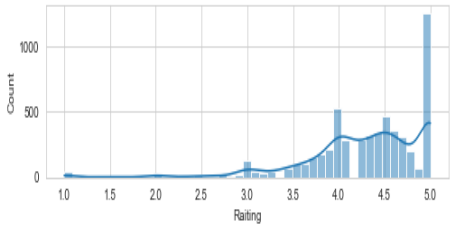
در این دیتاست هیچ missing value ای وجود ندارد. حال فیچر های مورد نظر را بررسی میکنیم.

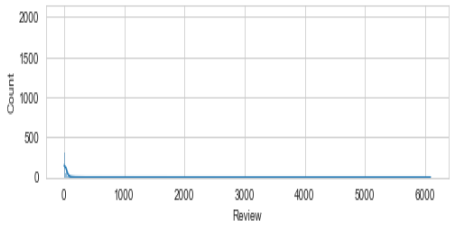
- Reviews and Ratings: به نظر میرسد نام ستون Reviews با Ratings عوض شده است. به همین انها را اصلاح میکنیم و مقدار هر یک را به صورت عددی به دست میاوریم.
- Edition: این ستون که نشان دهنده تاریخ چاپ است را به دو ستون مجزا تقسیم کردیم یک ستون Edtion date و دیگری ستون Edition type که به ترتیب شامل تاریخ و نوع کتاب چاپ شده است.
- به نظر میرسد فیچر های نام کتاب و ژانر و همچنین خلاصه کتاب تاثیری در قیمت نداشته باشند به همین دلیل ان ها را حذف کردیم.
- BookCategory:

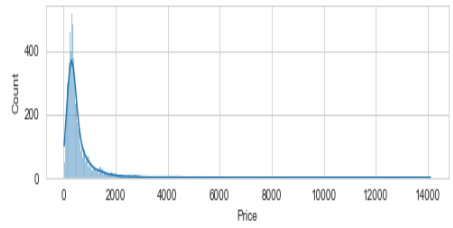
در این قسمت تعداد کتاب ها در هر کتگوری را بدست اوردیم و به نظر میرسد کتاب های بخش Computing,internet and digital media دارای بالاترین قیمت هستند.



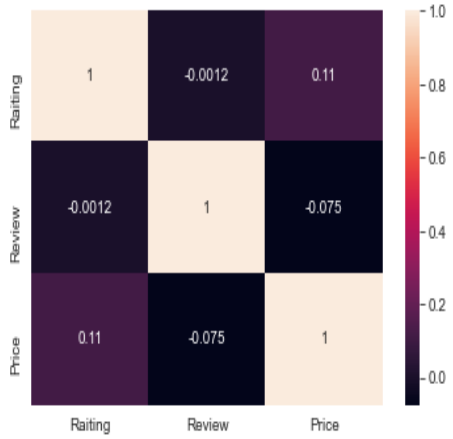
- در این قسمت هیستوگرام فیچرهای عددی را کشیدیم که نشان دهنده وجود چولگی است.

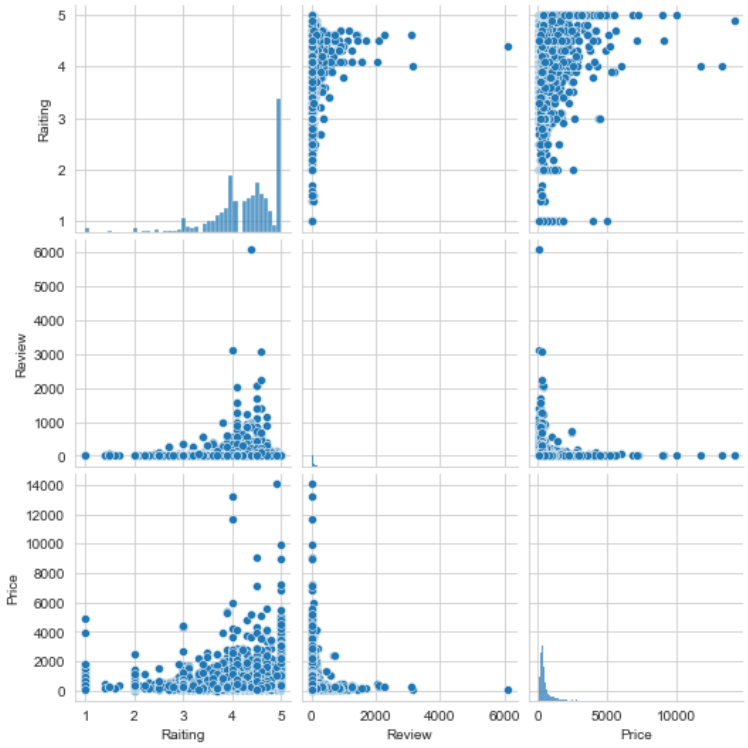






- با توجه به شکل هاي زير همبستگي بين فيچرهاي عددي مشاهده نميشود.





- Encoding:

سر انجام داده هاي train و test را بر اساس label encoding ، encode كرديم زيرا در هر فيچر كلاس هاي زيادي وجو داشت و استفاده از one hot encoding و يا get dummise به ديتاست

sparsity به دیتاست sparsity اضافه میکرد .

- Scaling:

بعد از ان از روش هاي scale کردن استفاده کردیم. براي فیچرها از standard scaler استفاده کردیم و براي ستون target از boxcox استفده کردیم که به شدت خطا را کاهش داد.

- Random forest regressor and MSE

و سپس از روش random forest regressor استفاده کردیم و مدل را اموزش دادیم. random forest regressor یک الگوریتم یادگیری نظارت شده است که از یک روش یادگیری گروهی برای رگرسیون در یادگیری ماشین استفاده می کند. درختان در جنگل های تصادفی به صورت موازی حرکت می کنند، به این معنی که هیچ تعاملی بین این درختان در هنگام ساخت درخت وجود ندارد. سرانجام مقدار خطاي mse برابر با0.01 شد.