

گزارش تمرین mobile-price-classification

از 2 تا 10:

روش Univariate

برای انتخاب ویژگی‌هایی که قوی‌ترین رابطه را با متغیر خروجی دارند، می‌توان از آزمون‌های آماری استفاده کرد.

کتابخانه scikit-learn کلاس SelectKBest را ارائه می‌دهد که می‌تواند با مجموعه‌ای از آزمون‌های آماری مختلف برای انتخاب تعداد خاصی از ویژگی‌ها استفاده شود.

از آزمون آماری χ^2 (chi-squared) برای ویژگی‌های غیرمنفی استفاده می‌کنیم تا K ($k=10$) از بهترین ویژگی‌ها را از مجموعه داده‌های پیش‌بینی محدوده قیمت موبایل انتخاب کند.

با استفاده از نمره بالا می‌توان نتیجه گرفت که "ram" مهم‌ترین ویژگی در بین همه ویژگی‌ها است.

11 تا 25:

Feature Importance

با استفاده از ویژگی feature importance مدل می‌توانیم اهمیت ویژگی هر یک از ویژگی‌های مجموعه داده خود را به‌دست بیاوریم.

اهمیت ویژگی به ما امتیازی برای هر یک از ویژگی‌های داده را می‌دهد، هر چه امتیاز بالاتر باشد، ویژگی نسبت به متغیر خروجی مهم‌تر یا مرتبط‌تر است.

feature importance، یک کلاس داخلی است که با طبقه‌بندی‌های مبتنی بر درخت ارائه می‌شود، اما در اینجا در این مثال از XGB Classifier برای استخراج 10 ویژگی برتر مجموعه داده استفاده می‌کنیم.

Correlation Matrix

همبستگی بیان می‌کند که چگونه ویژگی‌ها با یکدیگر یا متغیر هدف مرتبط هستند.

همبستگی می‌تواند مثبت باشد (افزایش یک مقدار ویژگی، مقدار متغیر هدف را افزایش می‌دهد) یا منفی (افزایش یک مقدار ویژگی، مقدار متغیر هدف را کاهش می‌دهد).

Heatmap تشخیص این‌که کدام ویژگی بیشتر به متغیر هدف مرتبط است را آسان می‌کند، این نقشه را برای ویژگی‌های همبسته با استفاده از کتابخانه seaborn ترسیم خواهیم کرد.

در بخشی دیگر از کد همچنین مدل مورد نظر پیاده‌سازی شده است.