نام ونام خانوادگی : سارا رضایی
شماره دانشجویی: ۹۸۲۲۲۰۴۳
عنوان : گزارش و تحلیل دیتاست اول تمرین ۲ درس مبانی یادگیری ماشین

مقدمه

داده های مورد بررسی در این تمرین داده های فروش تلفن های همراه شرکت های مختلف است. در این داده ها رابطه بین ویژگی های مختلف تلفن همراه و کلاس قیمت تلفن بررسی میشود.

در این تمرین قصد داریم با روش های مختلف فیچر های ورودی به مدل پیش بینی قیمت را انتخاب کنیم و نتایج پیش بینی مدل ها را مقایسه کنیم.

متد ها و مدل ها

از بلاک اول تا قبل از هدینگ تسک ۱ لود و بارگذاری داده ها از کگل انجام شده و shape و head و info دیتافریم برای مشاهده ی کلی داده ها چاپ شد. بررسی دیتافریم برای نداشتن داده null و outlier نیر در این قسمت انجام شده است .همچنین در این قسمت با کمک StandardScalar مقادیر دیتافریم استاندارد سازی شده اند و درنهایت در بلاک آخر دیتافریم در یک دیتافریم دیگر کپی شده تا دیتافریم بدون تغییری برای قسمت امتیازی موجود باشد.

تسک اول

انتخاب فیچر ها با روش forward selection

در روش forward selection ابتدا یک زیر مجموعه تهی از ویژگی‌ها ساخته می‌شود. سپس در هر مرحله، ویژگی‌هایی که بهترین عملکرد را برای مدل یادگیری به ارمغان می‌آورند، به این زیر مجموعه اضافه می‌شوند.

در این قسمت برای سادگی پیاده سازی کلاس قیمت مدل ها از ۴ کلاس به ۲ کلاس ادغام شده کاهش یافته اند.

معیار توقف forward selection مقدار auc است که با یک مدل logistic regression محاسبه کردم.

در قسمت بعدی تابع forward_selection مشاهده میشود که در نهایت فیچر های ['ram', 'px_height' , 'battery_power', 'px_width'] انتخاب شده اند.
و auc در معیار توقف برابر 0.99 بوده است.

تسک دوم

train کردن یک مدل logistic regression با فیچرهای انتخاب شده در قسمت قبلی میباشد. که برای هر مدل معیار های precision، recall، f1-score گزارش شده است.

در این قسمت مدل با فیچر های تابع forward_selection معیار f1-score آن برابر 0.99 میباشد.

همچین یک مدل random forest نیز در این قسمت train شده است که با فیچر های تابع forward_selection معیار f1-score آن برابر 0.9725 میباشد.

تسک سوم

اعمال PCA (Principal Component Analysis)

یکی از کاربردهای اصلیِ PCA در عملیاتِ کاهشِ ویژگی (Dimensionality Reduction) است.

PCA همان‌طور که از نامش پیداست می‌تواند مولفه‌های اصلی را شناسایی کند و به ما کمک می‌کند تا به جای اینکه تمامیِ ویژگی‌ها را مورد بررسی قرار دهیم، یک سری ویژگی‌هایی را ارزشِ بیشتری دارند، تحلیل کنیم.

در واقع PCA آن ویژگی‌هایی را که ارزش بیشتری فراهم می‌کنند برای ما استخراج می‌کند.

در این قسمت بر روی فیچر ها با تعداد component برابر با تعداد فیچرهای انتخاب شده در forward selection است که برابر ۴ است pca را اعمال کردیم.

تسک چهارم

Model training

در این قسمت یک مدل logistic regression با فیچرهای انتخاب شده در قسمت قبلی (PCA) train شده است. که برای مدل معیار های precision، recall، f1-score گزارش شده است.

در این قسمت معیار f1-score برابر 0.5309 است. که شاهد کاهش قابل توجه دقت مدل نسبت به حالت قبلی که با forward selection بود می باشیم.

تسک ششم

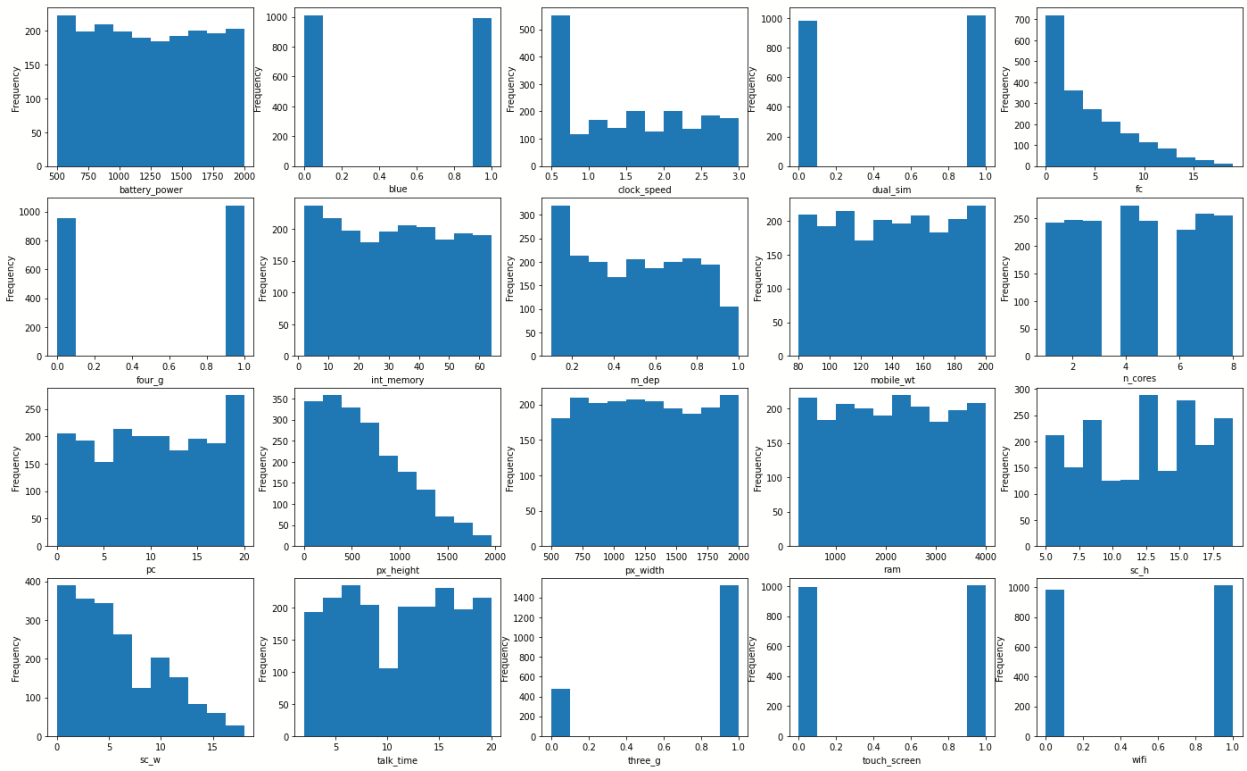
Feature Engineering

روند تعیین این که کدام ویژگی‌ها ممکن است در آموزش مدل مورد استفاده قرار بگیرند، و سپس تبدیل داده‌های خام موجود در منابع مختلف به آن نوع از ویژگی‌ها را مهندسی فیچر میگویند.

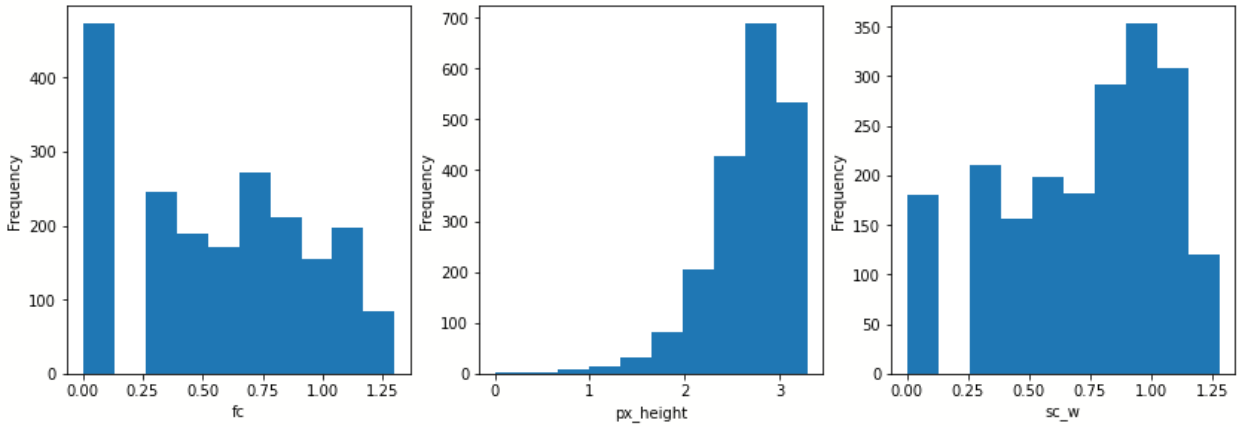
قسمت اول binning داده ها است. به طورکلی این روش یک روش همگام سازی داده هاست که در آن با بررسی همسایه ی هرداده سعی میشود داده را شبیه همسایه اش کند. و اگر داده ای تفاوت زیادی با همسایه اش داشت بدین معناست که داده نویزی است.
گام اول در این روش تعیین bin هاست که بدین منظور ابتدا مقادیر داده های به ترتیب صعودی یا نزولی sort میشوند و پس از آن مقادیر مرز bin ها را مشخص میکنیم.
در این مثال binning روی فیچر battery power انجام شده و با سه لیبل `'low','normal','high'` نام گذاری شده اند. که مرز این bin ها مقادیر `[0, 800, 1500, 3000]` میباشند.

قسمت دوم One-hot encoding است که توضیحات این روش در فایل کد آمده و روی داده های categorical اعمال شده است.

قسمت سوم log transformation است که توضیحات این تبدیل نیز در فایل کد آمده است و بر روی داده هایی که دچار skew distribution هستند اعمال شده است.



همانطور که در این تصویر مشاهده میشود فیچر های sc_w , px_height , fc فیچرهایی هستند که دارای skew distribution که برای اعمال log transformation انتخاب شده اند.



در این تصویر نتایج تبدیل لگاریتمی بر روی داده های تعیین شده نمایش داده شده اند.

و در نهایت در قسمت چهارم یک فیچر جدید با نام sc_area اضافه شده که مساحت گوشی ها است و از ضرب sc_w , sc_h بدست آمده است.

تسک هفتم

Train کردن svm بر روی داده های تغییر یافته در قسمت قبلی

svm (support vector machine) یا ماشین بردار پشتیبان یک الگوریتم نظارت‌شده یادگیری ماشین است که هم برای مسائل طبقه‌بندی و هم مسائل رگرسیون قابل استفاده است؛ با این حال از آن بیشتر در مسائل طبقه‌بندی استفاده می‌شود. در الگوریتمSVM ، هر نمونه داده را به عنوان یک نقطه در فضای-n بعدی روی نمودار پراکندگی داده‌ها ترسیم کرده (n تعداد ویژگی‌هایی است که یک نمونه داده دارد (و مقدار هر ویژگی مربوط به داده‌ها، یکی از مؤلفه‌های مختصات نقطه روی نمودار را مشخص می‌کند. سپس، با ترسیم یک خط راست، داده‌های مختلف و متمایز از یکدیگر را دسته‌بندی می‌کند.

این مدل بر روی ۵ حالت مختلف پیاده سازی شده است که نتایج آن به شرح زیر میباشد:

۱.بر روی دیتافریم اصلی svm.score برابر ۰,۹۶۵ است.

۲.بر روی دیتافریمی که فیچر battery_power در آن binning شده و سپس با one-hot-encoding کدگذاری شده است svm.score برابر ۰,۸۰۵ است.

۳.بر روی دیتافریمی که بر روی تعدادی از فیچرها log transformation انجام شده است svm.score برابر ۰,۹۲۲۵ است.

۴.بر روی دیتافریمی که فیچر جدید مساحت به آن اضافه شده است svm.score آن برابر ۰,۹۴۵ است.
۵.بر روی دیتافریمی که تمامی تغییرات ۴ مورد قبلی روی آن اعمال شده است svm.score آن برابر ۰,۸۱۵ است.

تسک هشتم و تسک نهم و تسک دهم سوالات تشریحی و تحلیلی بودند که در فایل کد به طور کامل به آن ها پرداخته شده است.

تسک های امتیازی

تسک اول

انتخاب فیچر ها با روش backward selection و train کردن مدل بر روی این فیچر ها

در این روش ابتدا تمامی ویژگی‌ها در زیر مجموعه حضور دارند. سپس در هر مرحله، بدترین ویژگی‌ها از زیر مجموعه حذف می‌شوند (ویژگی‌هایی که حذف آن‌ها، باعث ایجاد کمترین کاهش در عملکرد، دقت و کارایی روش یادگیری ماشین می‌شوند)

در این قسمت تعداد ۱۸ فیچر انتخاب شده است که به مدل logistic regression داده شده است و f1-score آن برابر ۰,۹۹ میباشد.