داده کاوی

تمرین شماره دو

بخش اول

ابراهیم شکیباپور

**درباره دیتاست**

باب شرکت تلفن همراه خود را راه اندازی کرده است. او می خواهد با شرکت های بزرگی مانند اپل، سامسونگ و غیره به مبارزه بپردازد.

او نمی داند چگونه قیمت موبایل هایی را که شرکتش تولید می کند تخمین بزند. در این بازار رقابتی تلفن همراه فقط حدس و گمان کافی نیست. برای حل این مشکل او داده های فروش تلفن های همراه شرکت های مختلف را جمع آوری می کند.

باب می‌خواهد رابطه‌ای بین ویژگی‌های یک تلفن همراه (به عنوان مثال: رم، حافظه داخلی و غیره) و قیمت فروش آن پیدا کند. اما او در یادگیری ماشین چندان خوب نیست. بنابراین او برای حل این مشکل به کمک شما نیاز دارد.

در این مشکل شما لازم نیست قیمت واقعی را پیش بینی کنید، بلکه یک محدوده قیمتی را که نشان می دهد قیمت چقدر بالا است.

**ستون های دیتاست**

**Battery power یا قدرت باتری :** کل انرژی که یک باتری می تواند در یک زمان ذخیره کند با واحد میلی آمپر ساعت (mAh) اندازه گیری می شود.

**Blue :** این که بلوتوث دارد یا خیر.

**Clock\_speed :** سرعتی که ریزپردازنده دستورالعمل ها را اجرا می کند.

**Dual sim :** از دو سیم کارت پشتیبانی می کند یا خیر.

**Fc :** مگاپیکسل های دوربین جلو.

**Four\_g :** فور جی دارد یا خیر.

**Int\_memory :** حافظه داخلی برحسب گیگابایت.

**M\_dep :** ضخامت موبایل به سانتی متر.

**Mobile\_wt :** عرض تلفن همراه.

**N\_cores : تعداد هسته های پردازشگر.**

**Pc : مگاپیکسل های دوربین اصلی.**

**Px\_height : ارتفاع رزولوشن پیکسل.**

**Px\_width : عرض رزولوشن پیکسل.**

**Ram : حافظه دسترسی تصادفی در مگا بایت.**

**Sc\_h : ارتفاع صفحه موبایل به سانتی متر.**

**Sc\_w : عرض صفحه موبایل به سانتی متر.**

**Talk\_time : طولانی ترین زمانی که یک بار شارژ باتری در زمانی که شما هستید دوام می آورد.**

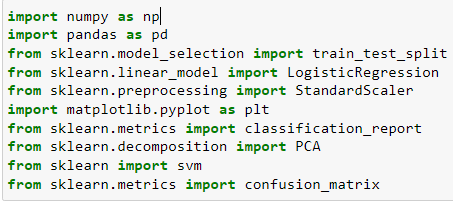
**Three\_g : تری جی دارد یا خیر.**

**Touch\_screen : صفحه لمسی دارد یا خیر.**

**Wifi : وای فای دارد یا خیر.**

**Price\_range‌ : این متغیر هدف با مقدار 0 (هزینه کم)، 1 (هزینه متوسط)، 2 (هزینه بالا) و 3 (هزینه بسیار بالا) است.**

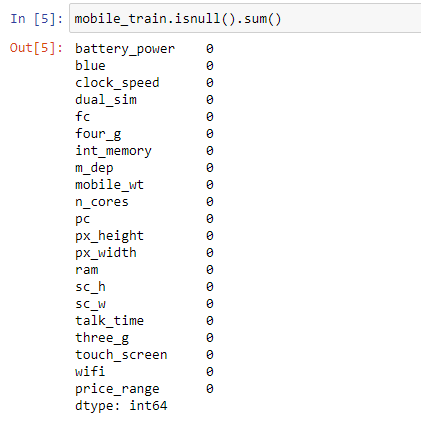
**ایمپورت کردن پکیج های مورد نیاز و خواندن دیتاست آموزش**





**بررسی دیتافریم**

دیتافریم کامل می باشد و فقدان داده در هیچ یک از ستون ها نداریم.

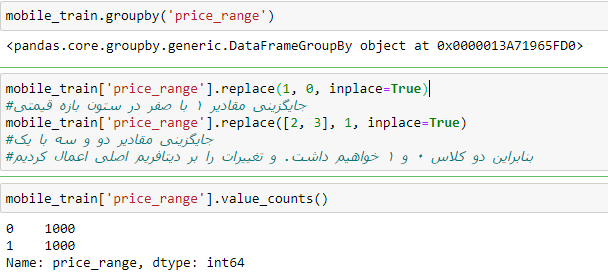


**۱ – پیاده سازی روش انتخاب پیشرو**

برای سادگی پیاده سازی, دیتافریم را براساس ستون بازه قیمتی گروه بندی می کنیم

دو کلاس دو و سه را که هزینه بالا و هزینه بسیار بالا بودند در یک کلاس گران یا بالا ادغام می کنیم : کلاس ۱

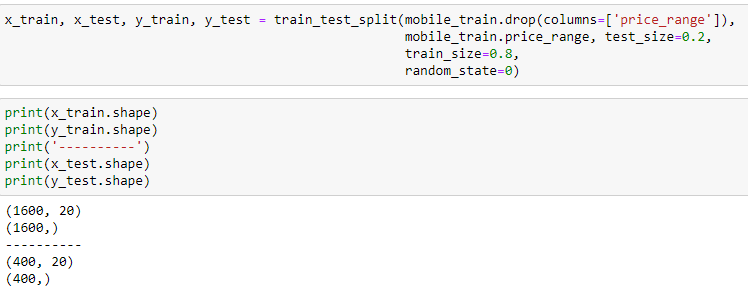
همچنین دو کلاس هزینه کم و هزینه متوسط را در یک کلاس ارزان یا پایین ادغام می کنیم: کلاس ۰



**تقسیم سازی داده به آموزش و تست**

تقسیم بندی دیتافریم به زیرمجموعه های آموزش و تست تصادفی

هشتاد درصد آموزش, بیست درصد تست و خروجی ها و لیبل های متناسب با هریک

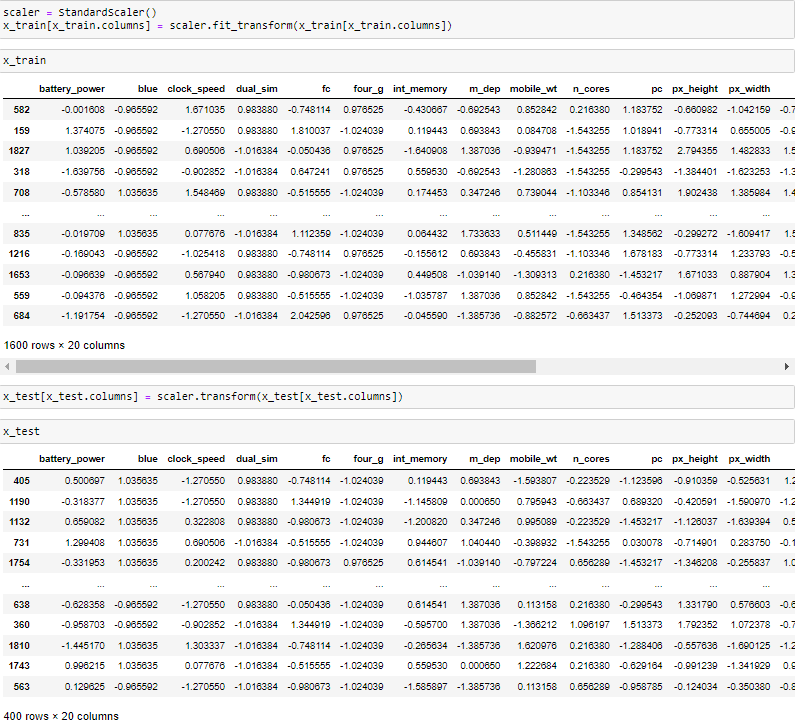


**تعریف تابع انتخاب پیشرو**

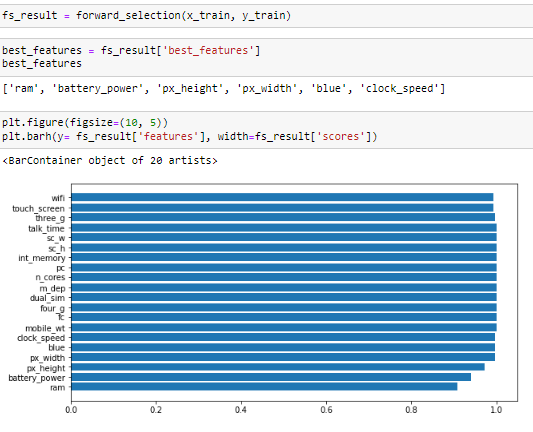
معیار انتخاب ویژگی جدید در هر مرحله هم AUC یا ناحیه زیر منحنی عملیاتی گیرنده می باشد. که البته در اینجا مجاز به استفاده از پکیج نمی باشیم ,به هر حال تعریف تابع انتخاب پیشرو در زیر داریم:



انتخاب ویژگی پیشرو با ارزیابی هر یک از ویژگیها شروع می شود و مواردی را انتخاب می کند که به بهترین عملکرد مدل الگوریتم انتخاب شده منجر شود, بهترین چیست؟!این کاملاً به معیارهای ارزیابی تعریف شده (ناحیه زیر منحنی عملیاتی گیرنده, دقت تخمین, جذر ميانگين مجذور خطا و ...) بستگی دارد .در مرحله بعد، تمام ترکیبات ممکن از ویژگی انتخاب شده و ویژگی بعدی مورد ارزیابی قرار می گیرند و ویژگی دوم انتخاب می شود و به همین ترتیب تا زمانی که تعداد مشخصه از پیش تعریف شده مورد نیاز انتخاب شود.



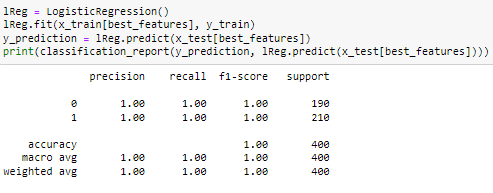
بعد از نرمال سازی آن را به تابع انتخاب پیشرو پاس می دهیم



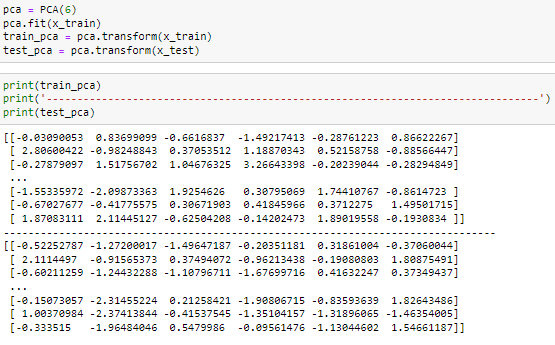
**۲ – اجرای مدل لجستیک بر روی داده ها**

‍بر اساس ویژگی های انتخاب شده از مرحله قبل مدل لجستیک را بر روی این ویژگی ها اجرا می کنیم. و گزارش معیار های زیر:

precision, recall, f1-score

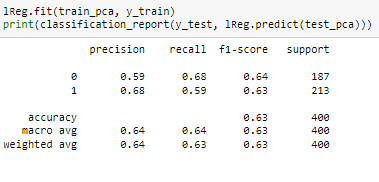


۳ - با استفاده از متد PCA می خواهیم ابعاد دیتاست اولیه خود را تغییر دهیم و به اندازه تعداد ویژگی های انتخاب شده در مرحله قبل برسانیم .آیا هنوز بیشتر اطلاعات دیتاست بزرگ را می توانیم با این دیتاست کوچک داشته باشیم؟



با استفاده از دیتاست تغییر یافته در مرحله قبل یک رگرسیون لجستیک را پیاده سازی می کنیم و گزارش معیارهای:

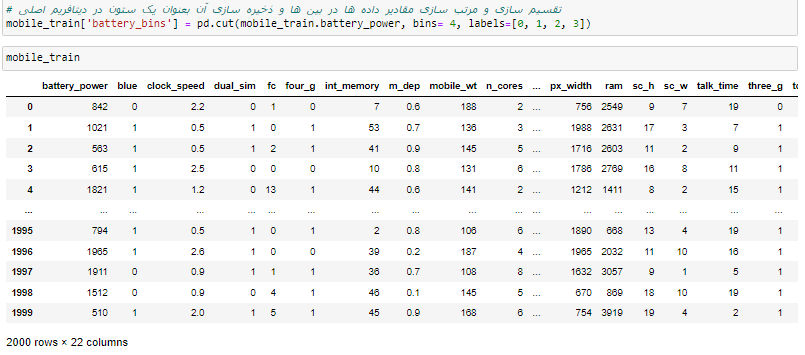
f1-score, recall, precision



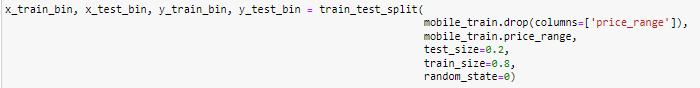
**۶ – مهندسی ویژگی**

## ۶ الف

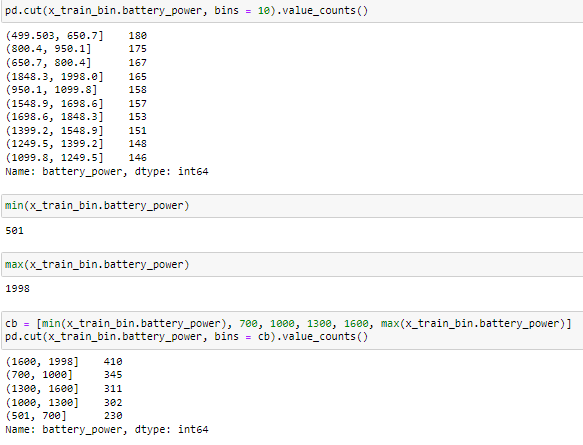
استفاده از روش بینینگ بر روی ویژگی قدرت باتری و استفاده از اندازه بین های مختلف



تقسیم سازی به داده آموزش و تست به نسبت هشتاد درصد به بیست درصد

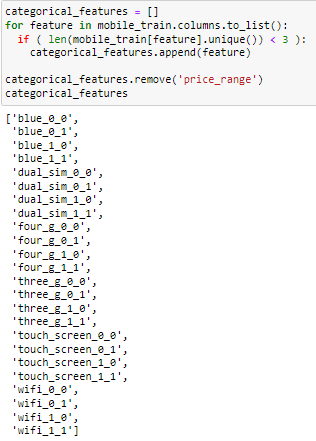
****

****

****

**۶ ب**

استخراج ویژگی های رده ای یا categorical و اعمال one hot encoding



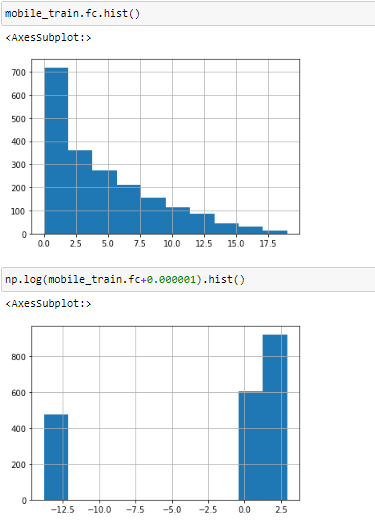


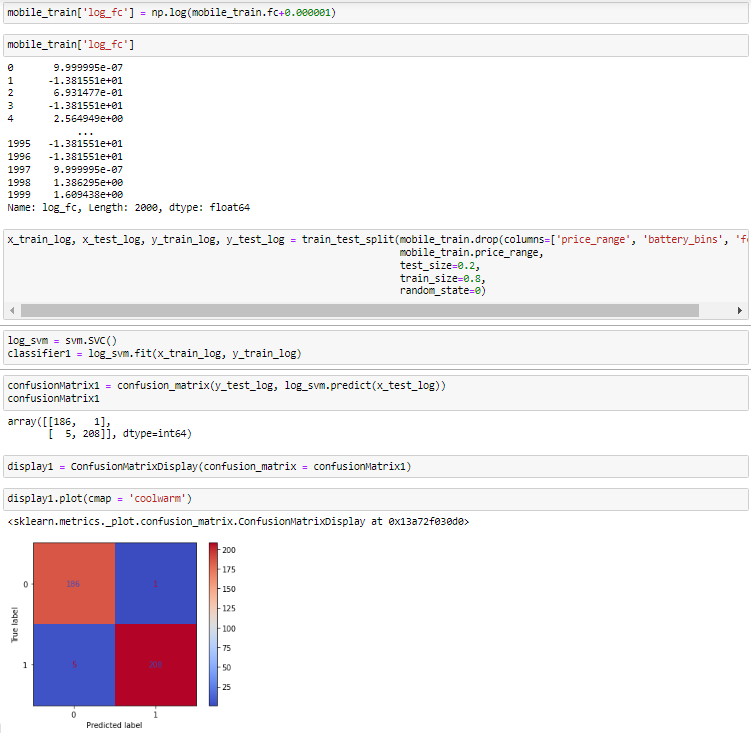
دلیل استفاده از : one hot encoding

بسیاری از مدل ها، امکان انجام پردازش مستقیم و بدون انجام «کدبندی» را روی داده های طبق ای ندارند بطور کل برای طبقه بندی داده های کتوگوریکال استفاده می شود. عملکرد آن به جز در مواردی که متغیر دسته ای مقادیر خیلی زیادی بگیرد، بسیار خوب است (معمولا پانزده) one hot encodingستون های دودویی جدیدی می سازد که هر یک مربوط به یکی از مقادیری هستند که متغیر به خود می گیرد. می توان گفت برای این است که بتوانیم روی داده ی عددی به گونه ای کار کنیم.

**۶ ج**

بررسی کنید آیا استفاده از تبدیل هایی از قبیل transform log و یا تبدیل نمایی در اینجا کاربرد دارد . به صورت کلی چرا از این دست تبدیلات بهره میبریم .

****

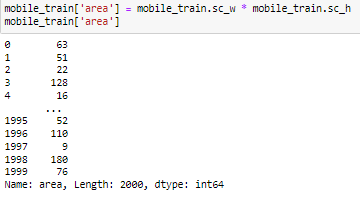
****

تبدیل لاگ را می توان برای کم کردن خمیدگی یا چولگی توزیع های با چولگی زیاد استفاده کرد

برای تفسیر پذیرتر کردن الگوهای موجود در داده ها

**۶ د**

یک ویژگی جدید به نام مساحت یا حجم گوشی



**۸**

بوت استرپینگ چیست؟

از روی یک دیتاست اورجینال b تا دیتاست بوت استرپ می سازیم. مثل این است که یه کیسه داشته باشیم و در آن دست کنیم و یک مورد را برداریم و به دیتاست بوت استرپ خود اضافه کنیم و دوباره آن را سر جای خود بگذاریم. پس امکان دارد آن مورد قبل دوباره تکرار شود.بی تا دیتاست می توان از روی دیتاست اصلی ساخت. بطور میانگین دیتاست بوت استرپ دو سوم دیتاهای اورجینال را دارند

دلیل استفاده از بوت استرپ؟

معمولا می خواهیم یک پارامتری را تخمین بزنیم پارامتری را که می خواهیم تخمین بزنیم از روی دیتاست اول محاسبه می کنیم سپس از روی دیتاست دوم محاسبه می کنیم و همین طور الی آخر برای هر بی تا دیتاست بوت استرپ.

بعد این ها را با هم جمع می کنیم و تقسیم بر تعداد می کنیم یعنی میانگین می گیریم. که واریانس کم می شود و تخمین به مقدار واقعی نزدیک می شود

تفاوت بوت استرپینگ با کراس ولیدیشن در این است که کراس ولیدیشن خود دیتاست اصلی را تکه می کند و به چندین دیتاست تقسیم می کند. در واقع مثل مثال کیسه ای که زدیم مقادیر را برنمی گرداند تا امکان تکرار امکان آنها وجود داشته باشد

سایز دیتاست هایی که با کراس ولیدیشن هم ساخته می شوند کوچکتر و کسری از دیتافریم اصلی است. ولی دیتاست های بوت استرپ سایزشان برابر با دیتاست اصلی است.

**۹**

5x2 Cross Validation همانطور که از اسمش مشخص هست این تست برای هر طبقه بند پنج بار 2-fold cross validation را اجرا می کند.

2-fold یعنی داده را در هر مرتبه به دو نیم تقسیم می کند نیمی برای آموزش و نیمی برای تست از اختلاف امتیاز برای محاسبه آماره تی استفاده می شود.

**۱۰**

الگوریتم های مختلف ساخت درخت تصمیم

به طور کلی چهار الگوریتم درخت تصمیم محبوب وجود دارد

ID3

CART (Classification and Regression Trees)

Chi-Square

و Reduction in Variance

**ID3**

آی دی تری از information gain استفاده می information gain برای ارزیابی این است که یک ویژگی چقد برای تقسیم بندی مناسب است در واقع در مسائل دسته بندی با استفاده از درخت تصمیم با دو چالش مواجه هستیم یکی انتخاب ویژگی مناسب و دیگر نقطه جداسازی یا همان .splitting pointاین اینفورمیشن گین تفاوت آنتروپی قبل و بعد از تقسیم بندی را محاسبه می کند که هرچقدر این مقدار بزرگتر باشد برای ما مناسب تر است . الگوریتم آی دی تری برای تصمیم گیری در مورد نقاط جداسازی از این information gainاستفاده می کند.

**CART**

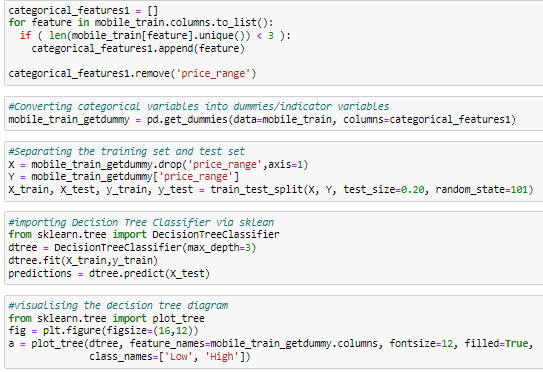
دیگر الگوریتم ساخت درختCART یا درخت های طبقه بندی و رگرسیون می باشد

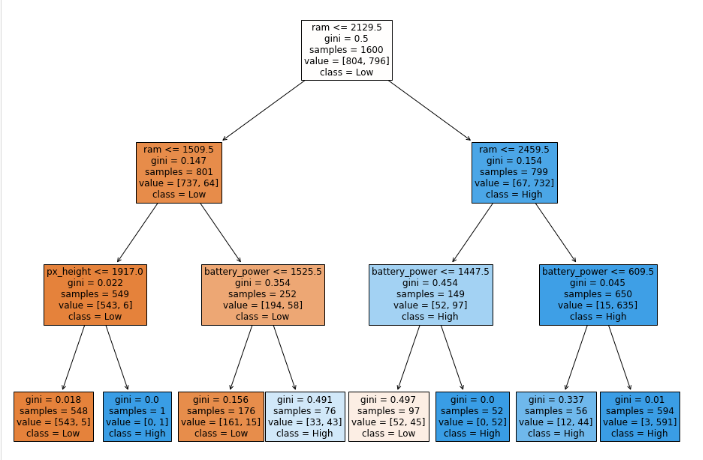
این الگوریتم از متد Gini (جینی ایندکس و جینی گین) برای ایجاد نقاط جداسازی (splitting point) استفاده می کند .جینی ایندکس احتمال تخصیص یک برچسب اشتباه به یک نمونه با انتخاب یک برچسب به طور تصادفی و همچنین برای اندازه گیری اهمیت ویژگی در یک درخت استفاده می شود

به طور کلی تفاوت الگوریتم های مختلف ساخت درخت در انتخاب نقطه جداسازی می باشد.

**۱۱**

**درخت تصمیم**

****

****

**۱۲**

**عمق درخت و تعداد نمونه های موجود و تعدد گره ها می تواند باعث خلوص گره ها شود. (node purity)**

**۱۳**

گاهی ممکن است ساخت درخت کامل روی دیتاست آموزش تخمین خوبی ایجاد کند ولی روی داده تست خوب کار نکند و بیش برازش یا overfitشود

آنقدر جداسازی را ادامه می دهیم که در هرشاخه یا جعبه مثلا پنج داده داشته باشیم می توانیم یک محک توقف قرار دهیم مثلا آنقدر تقسیم بندی بکنیم که در هر ناحیه مثلا یک داده داشته باشیم

در این حالت RSS صفر می شود ولی اگر تعداد داده ها در هر ناحیه خیلی کوچک شود آر اس اس نیز خیلی کوچک می شود و به عبارتی می گوییم برای این داده های آموزش خیلی خوب فیت و آموزش داده شده

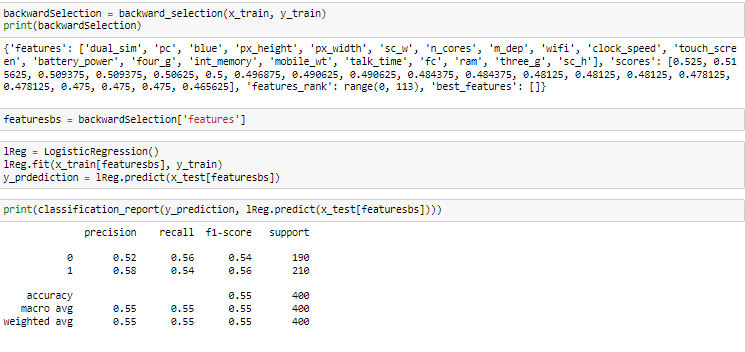
به دلیل جداسازی زیاد overfitting یا بیش برازش به وجود می آید پس کم کردن این جداسازی ها خوب می باشد. پس بدین منظور درخت را هرس می کنیم اگر درخت کوچکتری داشته باشیم که تعداد جداسازی های کمتری نسبت به آن داشته باشیم واریانس را پایین می آورد و بایاس نیز تا حد معقولی پایین می آید و tradeoff بین این دو حفظ می شود.

**۱۴**

در نمودار بایاس واریانس آن جایی که دو خط واریانس و بایاس هم دیگر را قطع می کنند تریدآف بین این دو می باشد و میزان مناسب میانگین خطاست

‌**Backward Selection**

****

****

## Statistical significance tests[¶](http://localhost:8888/notebooks/Desktop/project2/1/price.ipynb#Statistical-significance-tests)

مدل ها معمولاً با استفاده از روش های نمونه گیری مجدد(resampling) مانند اعتبارسنجی متقاطع کی فولد(k-fold cross validation) که در آن میانگین پارامترها محاسبه و به طور مستقیم مقایسه می شوند، ارزیابی می شوند

اگرچه ساده است، اما این رویکرد می تواند گمراه کننده باشد و منجر به نتایج و تصورات اشتباه شود، زیرا تشخیص اینکه تفاوت بین میانگین پارامترها واقعی است یا نتیجه یک تصادف آماری است دشوار است

**statistical significance tests** برای حل همین مشکل طراحی شده اند با این فرض طراحی شده اند که از توزیع یکسان گرفته شده اند. اگر این فرض یا فرضیه صفر رد شود، نشان می دهد که تفاوت از نظر آماری معنی دار است

اگرچه آزمون فرضیه های آماری بی اشتباه نیست، اما می تواند اعتماد شما را در تفسیر و ارائه نتایج در طول انتخاب مدل بهبود بخشد

## Matthews Correlation Coefficient

به عنوان معیاری برای کیفیت طبقه بندی های باینری (دو کلاسه) استفاده می شود اگر بیشتر داده ها در امتداد سلول های مورب قرار گیرند، دو متغیر باینری به طور مثبت مرتبط در نظر گرفته می شوند. در مقابل، اگر بیشتر داده ها از حالت مورب بودن خارج شوند، دو متغیر باینری به طور منفی مرتبط در نظر گرفته می شوند.