**نام و نام خانوادگی: غزل دانایی**

**شماره دانشجویی: 97222034**

**تاریخ تحویل: اریبهشت 1401**

دیتاست شماره 1:

سوال 2,1 -

بعد از انجام پیش پردازش های لازم و تغییر لیبل کلاس ها، تابعی برای محاسبه مقیاسauc برای هر مجموعه از ویژگی ها تعریف شد به این صورت که ویژگی کهauc بیشتری داشته باشد را به مجموعه ویژگی های منتخب اضافه می کنیم.

سپس مدل پیش بینی را روی داده ها با لیست ویژگی های بدست آمده آموزش می دهیم تا نهایتا روی مجموعه داده تست پیش بینی انجام بدهد. نتایج بدست آمده از این روش انتخاب ویژگی به صورت زیر می باشد:

ماتریس confusion به صورت :

239 25

31 246

می باشد بدین معنی که 25 داده در کلاس 0 بوده اند که به اشتباه در کلاس 1 پیش بینی شده اند(FP) و 31 مورد هم به به صورت برعکس این مورد استباه پیش بینی شده اند. (FN)

Recall = TruePositives / (TruePositives + FalseNegatives)=246/(246+31)=0.88

Precision = TruePositives / (TruePositives + FalsePositives)=246/(246+25)= 0.90

F1 = 2* Precision * Recall/( Recall+ Precision) = 1.584/1.78 = 0.88

در حالت بعدی برای انتخاب ویژگی ها از پکیج استفاده شد و بعد از آن به وسیله مدل پیش بینی انجام شد که با توجه به نداشتن FN,FP در این حالت انتخاب ویژگی نتیجه بهتری دارد و روش موثرتری می باشد.

Recall = 264/(264+0)=1

Precision 264/(264+0)=1

F1 = 2/2= 1

سوال 3 و 4-

نام و نام خانوادگی: غزل دانایی

شماره دانشجویی: 97222034

تاریخ تحویل: اریبهشت 1401

تغییر دیتاست با کمک PCA انجام شد و 15 تا ویژگی استخراج شد.سپس با استفاده از رگرسیون لجیستیک پیش بینی روی ویژگی هدف انجام شد نتایج به صورت confusion matrix زیر می باشد:

262 2

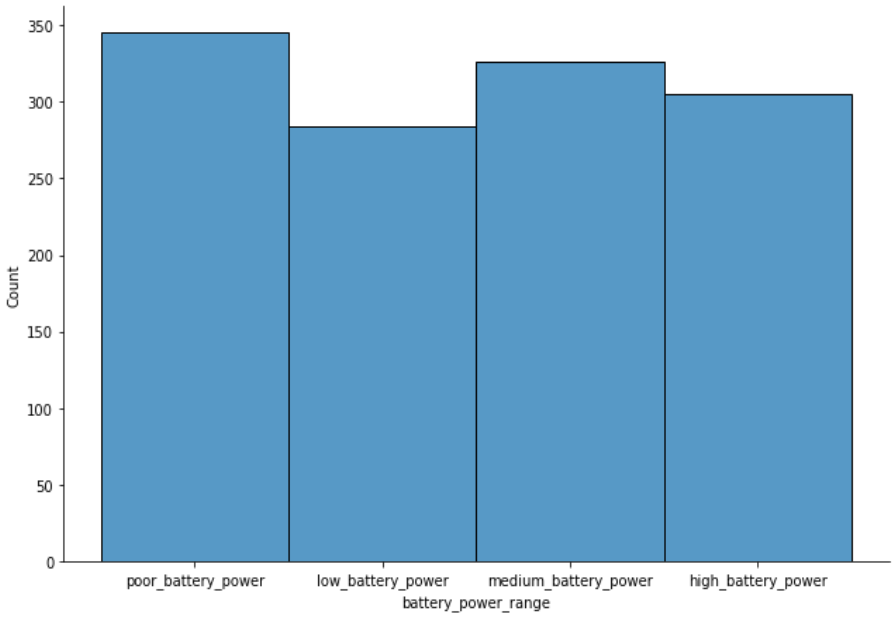
6 271

Recall = TruePositives / (TruePositives + FalseNegatives)=262/(262+6)= 0.97

Precision = TruePositives / (TruePositives + FalsePositives)=262/(262+2)= 0.99

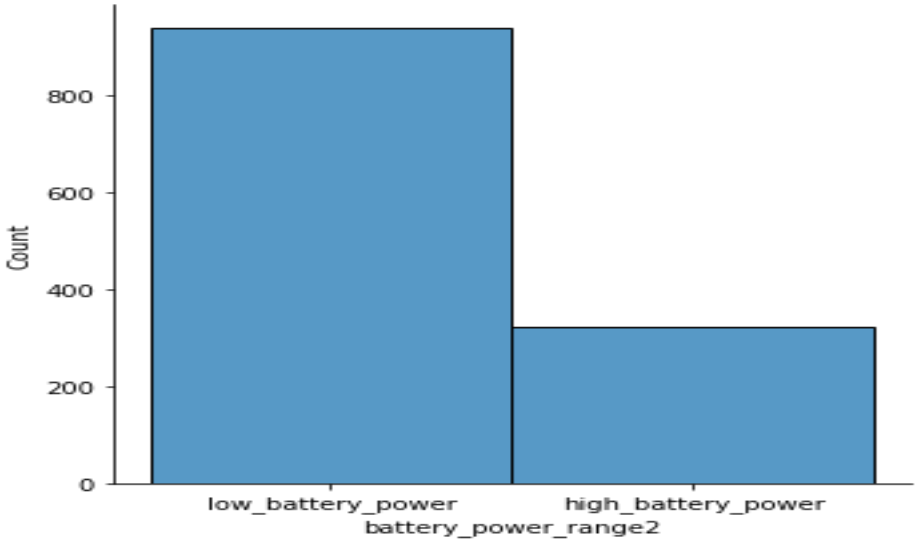
F1 = 2* Precision * Recall/( Recall+ Precision) = 2*0.99*0.97/1.96= 0.97

سوال6-قسمت الف)

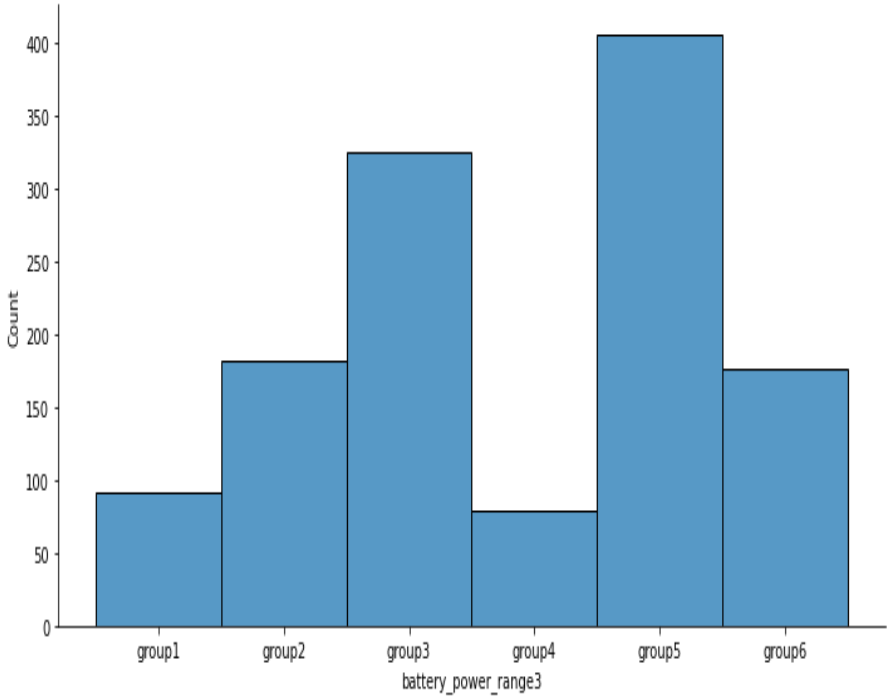



Poor battery power includes 501 to 874
Low battery power includes 875 to 1248
medium battery power includes 1249 to 1622
high battery power includes 1623 to 1998

نام و نام خانوادگی: غزل دانایی
شماره دانشجویی: 97222034
تاریخ تحویل: اریبهشت 1401




low battery power includes 501 to 1599
high battery power includes 1600 to 1998




Group 1 includes 501 to 599
Group 2 includes 600 to 799
Group 3 includes 800 to 1199

نام و نام خانوادگی: غزل دانایی
شماره دانشجویی: 97222034
تاریخ تحویل: اریبهشت 1401

```
Group 4  includes 1200 to 1299
Group 5  includes 1300 to 1799
Group 6  includes 1800 to 1998
```

سوال6-قسمت ب)

با استفاده از onehot encoding نمایش داده های categorical بهتر انجام می شود. (more expresive)تعداد زیادی از الگوریتم های یادگیری ماشین توانایی کار با داده های categorical را ندارند و باید این داده ها به صورت عددی تبدیل شوند. بعد از آن نیز میتوانیم روی داده ای عددی بدست آمده rescaling انجام دهیم.

سوال6-قسمت ج)

زمانی که داده ها در کلاس ها به صورت نامتوازن قرار گرفته باشند(imbalanced)، مدل روی تعداد بیشتری داده با کلاس خاص آموزش می بیند. مثلا اگر تعداد داده موبایل با قدرت باتری زیر 800 بیشتر از سایر مقادیر باشد، مدل بعدا برای پیش بینی داده های با قدرت باتری بیشتر از 800 به مشکل برمیخورد و کمتر مقدار توان باتری را درست پیش بینی می کند.

بعد از انجام binning در ویژگی جدیدbattery power range3 توزیع داده ها بین گروه ها نرمال نیست و می توان برای بهتر کردن عملکرد مدل از log transform استفاده کرد.

سوال7- الف)

دراین قسمت زمانی که دقیقا بعد از onehot encoding مدل را روی داده ها آموزش میدهیم و روی داده تست ارزیابی می کنیم، ماتریس confusion به صورت زیر می باشد:

263 1

2 275

سوال7-ب) نتایج مدل بعد از گرفتن لگاریتم

264 0

225 52

**نام و نام خانوادگی: غزل دانایی**

**شماره دانشجویی: 97222034**

**تاریخ تحویل: اریبهشت 1401**

سوال7-ج)نتیج مدل بعد از اضافه کردن فیچر ساخته شده:

264 0

224 53

قسمت د)روی همه حالات با نتایج:

263 1

2 275

سوال8-

Bootstrapping و cross validation هر دو از روش های resampling هستند. در این روش ها مجموعه داده اصلی به چندین دیتاست تقسیم می شود. در این روش ها دوباره یک مدل را روی نمونه های بدست آمده از مجموعه داده آموزشی فیت می کنیم تا اطلاعات دیگری راجع به مدل فیت شده بدست آوریم.

در K cross validation مراحل به صورت زیر می باشد:

1. دیتاست را به صورت تصادفی پخش می کنیم (shuffle)

2. دیتاست را به K گروه تقسیم می کنیم.

3. برای هر گروه یکتا ابتدا آن را به صورت دیتای تست و بقیه گروهها را به عنوان داده آموزشی در نظر می گیریم. مدل را روی داده های آموزشی فیت می کنیم و روی داده تست ارزیابی می کنیم. امتیاز ارزیابی را نگه می داریم و مدل را کنار می گذاریم.

4. در آخر به وسیله میانگین امتیاز ها در هر نمونه، کارایی مدل را بدست می آوریم.

- با bootstrapping تخمینی برای مقادیری از جمعیت(معمولا standard error )انجام می دهیم با میانگین گرفتن تخمین ها روی چندین نمونه داده کوچک. در هر بازدید بعد از انتخاب آنهارا به دیتاست اصلی برمیگردانیم که این به ما شانس انتخاب دوباره آن ها را می دهد در این روش نوعی نمونه سازی تصادفی با جابه جایی را داریم. مراحل کار به صورت زیر می باشد:

1. تعداد نمونه ها bootstrap برای اجرا و نیز اندازه نمونه ها را تعیین می کنیم
2. برای هر نمونه bootstrap یک نمونه با جابه جایی و همان اندازه انتخاب شده تشکیل می دهیم. سپس امار و ارقامی که مد نظرمان می باشد را روی این نمونه محاسبه می کنیم.
3. در آخر میانگین امار های به دست آمده از هر نمونه را محاسبه می کنیم.

- اگر از این روش برای تخمین کارایی مدل استفاده کنیم بعد از تشکیل هر نمونه میتوانیم مدل را روی آن آموزش دهیم (فیت کنیم) کارایی را در هر نمونه بسنجیم و بعد در آخر میانگین آن ها را حساب کنیم.
- از bootstrap برای محاسبه confidence intervals به صورت تقریبی نیز می توان استفاده کرد.

سوال 9-

در 5*2 cross validation در واقع 5 مرتبه cross validation با تعداد 2 گروه انجام می شود. این تست از McNemar's test قوی تر است . انتخاب این که کدام تست بهتر است بستگی به هزینه محاسباتی الگوریتم دارد. برای الگوریتم هایی که فقط یک بار می توانند اجرا شوند McNemar's test مناسب است اما برای الگوریتم هایی که از نظر محاسباتی سبک هستند و می توانند 10 مرتبه اجرا شوند 5*2 cv بکار برده می شود چون که قوی تر است و به صورت مستقیم واریانس را بنابر دیتای آموزشی محاسبه می کند.

سوال 10-

خیر نمی توان گفت که همواره با استفاده از این نمودار ها می توان مرتبه مناسب مدل را برای فیت شدن روی دیتای تست پیدا کرد چون در مواردی فرمول دقیق تابعی که آن را پیش بینی می کنیم مشخص نیست و نمی توان مقدار bias را محاسبه کرد.

نام و نام خانوادگی: غزل دانایی
شماره دانشجویی: 97222034
تاریخ تحویل: اریبهشت 1401

## تسک امتیازی –

### سوال2–

مدل ها معمولا با کمک روش های resampling مثل k-fold cross-validation بررسی می شوند که درآن میانگین امتیازات کارایی مدل ها محاسبه می شودو به طور مستقیم مقایسه می شوند اما این روش می تواند گمراه کننده باشد زیرا که مطمین نیستیم که این تفاوت در اعداد میانگین امتیاز ها واقعی است یا نتیجه شانسی بوده است.
بنابراین تست های آماری طراحی شدند تا مقدار likelihood در نمونه های امتیاز ها را اندازه گیری کنند با این فرض که از یک توزیع پیروی می کنند.اگر فرض صفر ما مبنی بر این که این میانگین ها در واقعیت متفاوت نیستند رد بشود میتوانیم نتیجه بگیریم که یک مدل واقعا از دیگری عملکرد بهتری دارد.مثلا یک تستی به نام McNemar's test معرفی شد تا به کمک آن برای الگوریتم های سنگین از نظر محاسباتی بتوانیم مقایسه انجام دهیم. مثلا در اینجا درست به معنی این است که می توانی فرض صفر این که دو تا کلاسیفایر خطاهای متفاوتی دارند را رد کنیم و غلط یعنی این که فرض صفر رد نمی شود و خطاهای هر دو مشابه یکدیگر می باشد.

TABLE 5. TEST RESULT COMPARISON OF TWO TESTS

| | McNemar | 5x2cv |
|---|---|---|
| XGB vs. LR | False | False |
| XGB vs. RF | **True** | False |
| XGB vs. SVM | **True** | False |
| LR vs. RF | False | False |
| LR vs. SVM | False | False |
| RF vs. SVM | **True** | **True** |

### سوال3–

Matthews Correlation Coefficient یک ابزار آماری برای بررسی مدل ها می باشد و به این صورت کار می کند که تفاوت بین مقادیر واقعی و پیش بینی شده را اندازه گیری میکند . در این روش که

مشابه chi-square می باشد به بیان دیگر ماتریس خلاصه سازی می شود. مقادیر این مقیاس بین 0 و 1 قرار می گیرد. 1 بهترین حالت بین مقادیر پیش بینی شده و واقعی است و 0 نیز به معنای این است که پیش بینی ها تصادفی بوده است.

میتوان گفت MCC به ما در فهمیدن ناکارآمدی کلاسیفایر در پیش بینی مخصوصا کلاس منفی ها کمک می کند.

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

نام و نام خانوادگی: غزل دانایی

شماره دانشجویی: 97222034

تاریخ تحویل: اریبهشت 1401