تقنيات تعلم الآلة

Machine Learning Techniques

1446/12/26

إشراف الدكتور عصام سلمان

إعداد:

محمد إحسان سرور C1 286368

قاسم أكرم الجبر C1 277434

حسين بهجت خضور C1 290482

مقدمة

نهدف في هذه الوظيفة إلى تحليل مجموعة البيانات المقدمة وبناء نموذج لمساعدة البنك في قبول ورفض الطلبات المقدمة إليه. حيث تحوي مجموعة البيانات المتوفرة على العديد من السمات المفيدة مثل العمر، عدد أفراد العائلة، مقدار القرض، إلخ. حيث سنستفيد من هذه السمات لبناء مصنف لاتخاذ القرار المناسب (رفض أو قبول الطلب المقدم)، حيث يحوي الطلب المقدم على معلومات القرض والمتقدم (السمات الموجودة في مجوعة البيانات) التي نقدمها للمصنف لاتخاذ القرار المناسب. وقمنا في هذه الوظيفة بالقيام بعملية تحليل البيانات ومعالجتها ومن ثم تدريب النموذج ومناقشة النتائج. كما قمنا ايضاً باستخدام الموقع colab.reseach.google.com الذي قمنا بتنفيذ الأكواد عليه.

بالإضافة إلى ذلك استخدمنا الأداة git والموقع github.com لرفع العمل الذي قمنا به وتحقيق التشاركية والتعاون في إنجاز العمل وإدارة الكود.

رابط الملف الذي عملنا عليه:

https://colab.research.google.com/drive/1FIymUq_ewEzv_Ka5_oLnJ5Nmuujy3sV_?usp=sharing

رابط الموقع: https://husenkh.pythonanywhere.com

رابط المستودع على github.com: https://github.com/husenKhaddour/loan-prediction

المعالجة المسبقة للبيانات

قمنا هنا أولاً بتحميل مجموعة البيانات والتي هي بصيغة CSV، ومن ثم استعرضنا الحقول (السمات) وتعرفنا عليها، وقمنا بتحليلها حيث يبين الجدول أدناه ملخص عملية التحليل لهذه السمات (مجالها وعدد القيم غير الموجودة وتوزعها).

العمود	الوصف	مجال القيم	توزع القيم	القيم الفارغة
Loan ID	معرف القرض	#	#	0
Gender	جنس المتقدم	{Male , Female}	18% F ,82% M	13
Dependents	عدد أفراد عائلة المتقدم	{0,1,2,+3}	[57%, 17%, 17%, 8%]	15
Education	مؤهلات المتقدم	{Graduated, Un Graduated}	[78% , 22%]	0
Property Area		{Semi Urban, Urban, Rural}	[39%, 31%, 30%]	0
Self Employed	هل المتقدم موظف	{No, Yes}	[81%, 19%]	32
Married	الوضع العائلي للمتقدم (متزوج أم لا)	{Yes , No }	[65%, 35%]	3
Loan Status	حالة القرض (مقبول أم مرفوض)	{eligible , not eligible}	[68%, 32%]	0
Credit History	تاريخ الائتمان (هل تقدم بطلبات قبل)	{Yes, No}	[73%, 27%]	50

Applicant Income	دخل المتقدم	[0, 72000]	mean: 4805, STD: 4910	0
Loan Amount	مقدار القرض	[108, 550]	mean : 136 , STD: 61	14
Co app. Income	دخل الكفيل	[0,24000]	mean: 1569, STD: 2334	0
Loan Amount Term	مدة القرض	[6,480]	mean: 34, STD:65	22

من أجل التعامل مع القيم الفارغة (غير الموجودة) وهو أمر مهم يجب التعامل معه في معظم مجموعات البيانات. حيث نكون أمام حلين إما إزالة القيم غير الموجودة أو استبدالها بقيم تتناسب مع توزع القيم الأخرى (المتوسط، القيمة الأكثر تكراراً، القيمة الأعلى). وفي حالتنا هذه ولأنّ عدد القيم الفارغة قليل، فقررنا استبدالها كما يلي: بالنسبة للقيم العددية Numerical Values فاستبدلناها بالمتوسط Mean لأنّ ذلك لا يغير من توزع البيانات في العمود. وبالنسبة للقيم الفئوية Categorical Values فاستبدلناها بالقيمة الأكثر تكراراً Mod. ومن أجل التعامل مع القيم النصية قمنا بترميزها أي بتحويلها من نصوص إلى أرقام (مثل Male ,female => 1,0).

ولأننا سنتعامل مع نماذج قد تكون حساسة لمجال القيم (مثل SVM) لذلك قمنا بتنظيمها Normalization أي مثلاً تحويل مجال الراتب من [100,10000] إلى المجال [0,1]، وهذه العملية مهمة لأن بعض النماذج حساسة لمجال القيم (أي قد تنحاز لمجال ما إذا كان أكبر من آخر). حيث يوضح الشكل أدناه البيانات بعد معالجتها:

Gender	Married	Dependents	Education	SelfEmployed	ApplicantIncome	CoapplicantIncome	LoanAmount	LoanAmountTerm	CreditHistory	PropertyArea	target
1	0	0	0	0	0.187257	-0.231114	0.104512	8	1	2	1
1	1	1	0	0	0.059339	0.033401	0.017251	8	1	0	0
1	1	0	0	1	-0.100608	-0.231114	-0.276588	8	1	2	1
1	1	0	1	0	-0.142742	0.182498	-0.020664	8	1	2	1
1	0	0	0	0	0.202514	-0.231114	0.078862	8	1	2	1

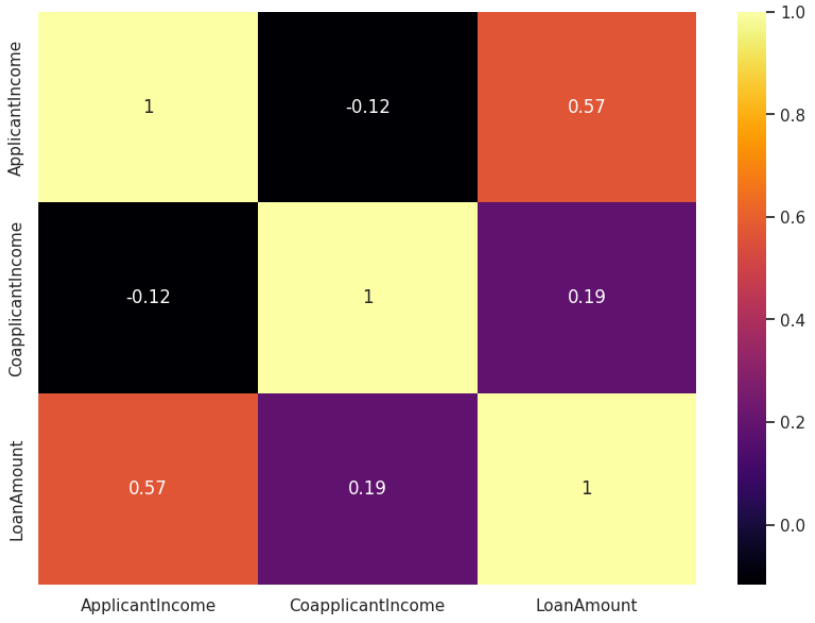
رسم توضيحي 1: عينة من مجموعة البيانات بعد القيام بعمليات المعالجة عليها.

تحليل البيانات واستكشافها EDA

لنبدأ بتحليل القيم العددية، أولاً نلاحظ أن مدة القرض هي فئوية تكون على سنة، سنتين، لغاية 8 سنوات، على عكس ما كنا نتوقع، لذلك لاحقاً تعاملنا معها كقيمة فئوية. أما بالنسبة لمقدار القرض فنلاحظ أنّ له توزيعاً طبيعياً، بينما دخل المتقدم له توزيع منحاز لليسار (القيم القليلة)، بينما نلاحظ أن هنالك عدداً من الكفلاء (15%) ليس لهم دخل، وأن توزيعه منحاز لليسار أيضاً.

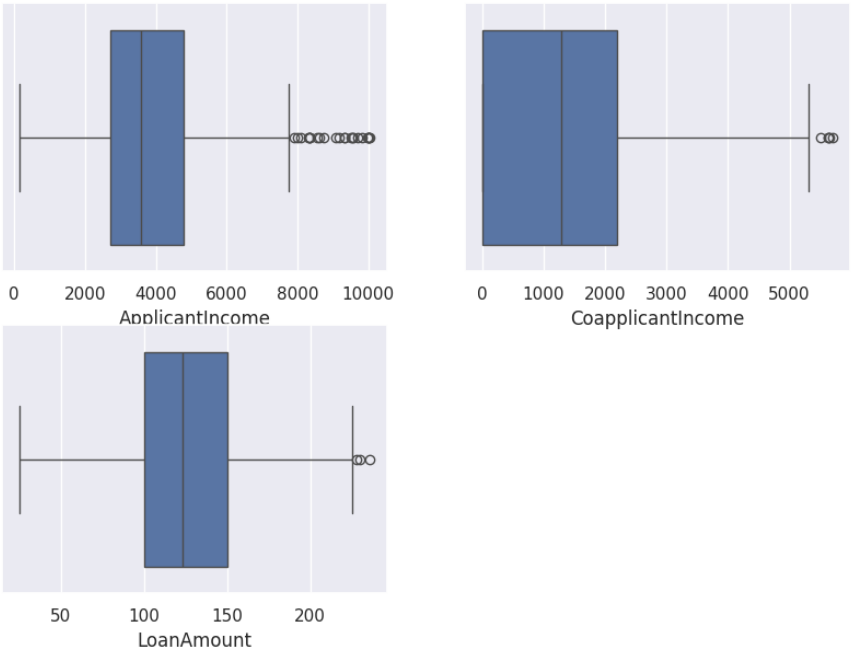
رسم توضيحي 2: توزيع القيم العددية.

وعند تحليلنا للقيم العديدة والترابط فيما بينها، ودراسة مصفوفة الترابط الموضحة أدناه، نلاحظ أن هنالك ترابط بين دخل المتقدم ومقدار القرض، وهذا شيء متوقع لأن البنوك تفرض تكافؤ بين الدخل ومقدار القرض، والأمر الآخر هو عدم وجود ترابط بين دخل المتقدم ودخل كفيله وهذا يعني أن البنك يقبل الكفيل بغض النظر عن تكافؤ دخله مع القرض وهو أمر غير متوقع.



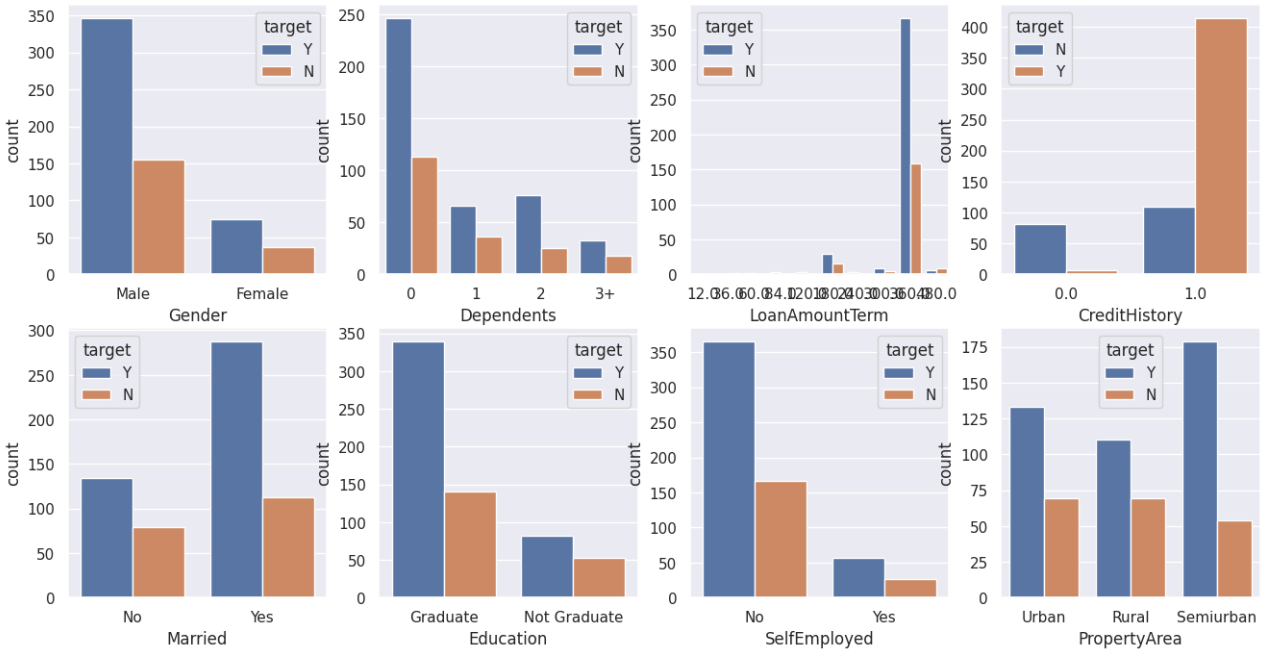
رسم توضيحي 3: مصفوفة الترابط Correlation Matrix.

وفي الشكل أدناه نرى مخططاً صندوقياً يوضح توزع القيم، ويظهر لنا النقاط الـ outliers. حيث بالنسبة لدخل المتقدم فنرى بأنها الأشخاص الذين لهم دخل مرتفع جدا جدا (أكثر من 8000، بينما وسطي دخل الأشخاص هو 4000)، وكذلك الأمر بالنسبة لدخل الكفيل.



رسم توضيحي 4: مخطط صندوقي لتوزع القيم العددية.

وعند دراسة الحقول الفئوية، نلاحظ أن معظم المتقدمين من الذكور (82%)، وأغلبهم يفون قروضهم، ويلاحظ أيضاً أن هنالك علاقة بين قبول القرض وتاريخ ائتمان المتقدم (أي معظم 70% الذي يتعاملون مع البنك في الماضي، هم أشخاص يتم قبولهم). كما يلاحظ بأنّ أغلب المتقدمين للقرض هم من الأشخاص الموظفين، والأشخاص الجامعيين. وكذلك الأمر بالنسبة لمتزوجين.



رسم توضيحي 5: توزع القيم الفئوية مع النظر لحالة القرض.

تعلم تصنيف طلبات القروض

قمنا بتقسيم مجموعة البيانات إلى مجموعة تدريب ومجموعة اختبار وفق النسبة 30/70. لم نضف مجموعة التحقق validation لضبط المعاملات الفوقية، لأننا استخدمنا خوارزمية البحث Grid Search حيث هي

تقوم بذلك (أي تقسم المجموعة التدريب إلى مجموعتين تدريب وتحقق) وتبحث ضمن مجالات المعاملات الممررة. كما هو موضح أدناه.

```
# تعريف موسطات البحث
parameters = {
    # أكبر عمق للشجرة
    'max_depth': [8,10,15,13,15],
    # أقل عدد عناصر لتقسيم العقدة
    'min_samples_split': [35,45,50,60],
    # أقل عدد عناصر لاعتبار عقدة ورقة
    'min_samples_leaf': [20,30,25,35],
    # معيار تقويم الشجرة
    'criterion': [ 'entropy']
}

DT_grid_search = GridSearchCV(DecisionTreeClassifier(criterion='entropy'), parameters,scoring='f1', cv=5)
```

وبعد تقسيم مجوعة البيانات إلى تدريب واختبار، قمنا بتدريب ثلاثة نماذج وهي SVM، Decision Trees، KNN. حيث حصلنا على نتائج متشابهة بين الـ SVM والـ Decision Trees

وكانت ملاءمتها هي: 83%، 83%، 78%.

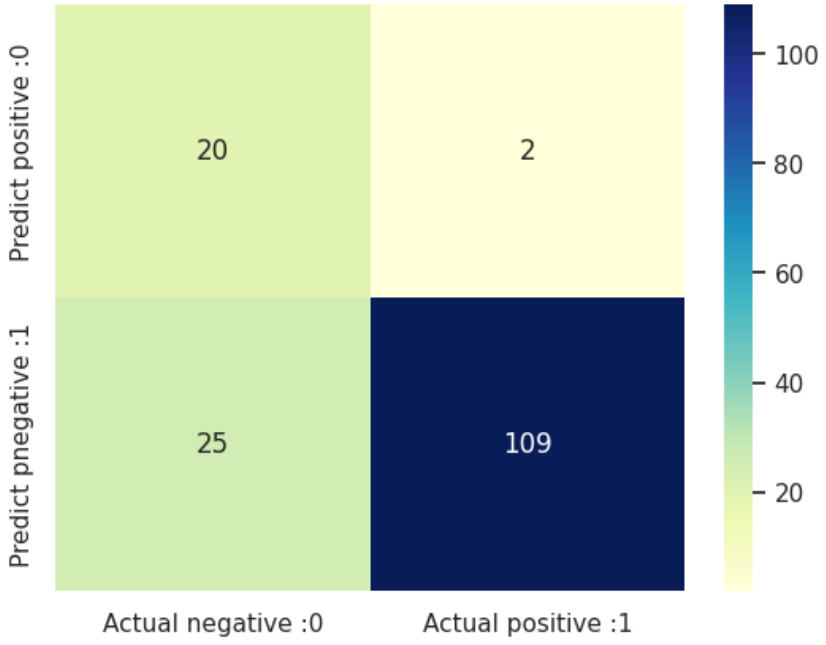
```
# تقرير التصنيف
print(classification_report(y_test, SVM_y_pred))
```

```
              precision    recall  f1-score   support

           0       0.91      0.44      0.60        45
           1       0.81      0.98      0.89       111

    accuracy                           0.83       156
   macro avg       0.86      0.71      0.74       156
weighted avg       0.84      0.83      0.81       156
```

رسم توضيحي 6: تقرير التصنيف.



رسم توضيحي 7: مصفوفة الـ Confiusion

نلاحظ من الشكل 7 أنّ النموذج قادر على تذكر الأشخاص الذين سيفون بالقروض أي لن يرفض (قليل جدا) شخص يمكن أن يفي بقرضه، وانّه قد يقبل أشخاص لن يفون بقروضهم.

	precision	recall	f1-score	support
0	0.72	0.40	0.51	45
1	0.79	0.94	0.86	111
accuracy			0.78	156
macro avg	0.76	0.67	0.69	156
weighted avg	0.77	0.78	0.76	156

رسم توضيحي 8: تقرير التصنيف لنموذج KNN.