تصنيف تعثر سداد القروض

باستخدام التعلم الآلي

إنجي غبيس
كلية الهندسة المعلوماتية-جامعة دمشق

باسل عامر
كلية الهندسة المعلوماتية-جامعة دمشق

خديجة حاجي
كلية الهندسة المعلوماتية-جامعة دمش

دانا كلش
كلية الهندسة المعلوماتية-جامعة دمشق

روبين حسو
كلية الهندسة المعلوماتية-جامعة دمشق

الملخص التجريدي

تعاني المؤسسات المالية من تحديات كبيرة في التنبؤ بتعثر سداد القروض بسبب عدم توازن البيانات ووجود قيم مفقودة أو شاذة. تهدف هذه الدراسة إلى بناء نموذج تنبؤي باستخدام تقنيات تعلم الآلة لتصنيف حالات التخلف عن السداد بدقة عالية. شملت الدراسة تحليلًا شاملًا لمجموعة بيانات حقيقية، مع تطبيق منهجيات متقدمة لمعالجة البيانات، مثل التشفير، والتعامل مع القيم المفقودة والشاذة، وهندسة السمات.

تمت مقارنة عدد من خوارزميات تعلم الآلة، منهاDecision Tree، XGBoost، LightGBM، CatBoost، Random Forest، بالإضافة إلى نماذج تجميع مثل Voting و Stacking .كما تم تقييم تأثير استراتيجيات إعادة التوازن (SMOTE, ADASYN, Class Weight, SMOTE+Tomek) على أداء النماذج.

أظهرت النتائج أن نماذج مثل LightGBM و Stackingتفوقت من حيث F1-score والدقة التنبؤية، بينما بقي نموذج Decision Tree مناسبًا كنقطة مرجعية بسيطة. توضح هذه الدراسة أهمية دمج تقنيات الضبط التلقائي للمعلمات (Optuna) ومعالجة البيانات بعناية للحصول على أداء تنبؤي موثوق وفعّال في مجال التنبؤ بتعثر القروض.

الكلمات المفتاحية: التنبؤ، بتخلف سداد القروض، بيانات غير متوازنة, تعلم الي .

I. المقدمة

تُعد مشكلة تعثر المقترضين عن سداد القروض من أبرز التحديات التي تواجه المؤسسات المالية، لما لها من تأثير مباشر على الأداء المالي واستقرار النظام الائتماني.

ومع تطور تقنيات تحليل البيانات، بات من الممكن الاستفادة من أدوات الذكاء الاصطناعي، خاصة خوارزميات التعلم الآلي، في بناء نماذج تنبؤية تساعد في الكشف المبكر عن حالات التعثر المحتملة. رغم ذلك، تواجه هذه النماذج تحديات عدة تتعلق بطبيعة البيانات المالية، أبرزها عدم توازن التوزيع بين حالات التعثر والسداد، ووجود قيم مفقودة أو متطرفة، فضلًا عن تعقيد العلاقات بين المتغيرات.

مما يتطلب اتباع منهجيات دقيقة في معالجة البيانات واختيار النماذج المناسبة لضمان دقة التنبؤ. تهدف هذه الدراسة إلى بناء نموذج تنبؤي يعتمد على تقنيات تعلم الآلة لتصنيف حالات التعثر، مع التركيز على معالجة التحديات المرتبطة بجودة البيانات وتوازنها. كما تستعرض الدراسة فعالية عدد من الخوارزميات الشائعة في هذا المجال، بهدف تقديم إطار يساعد المؤسسات المالية على تحسين قرارات الإقراض وتقييم المخاطر.

II. الدراسة المرجعية

العديد من الأبحاث تهتم بدراسة التنبؤات بتخلف سداد القروض لأهميته في المؤسسات المالية وتأثيرها على الربحية، في هذا القسم، يتم استعراض عدد من الدراسات التي ركزت على استخدام تقنيات مختلفة من تعلم الآلة لبناء نماذج فعالة للتنبؤ بتعثر السداد.

قدم Zhouو Wang .في (2012) [1] خوارزمية محسنة لخوارزمية الغابات العشوائية (Random Forests) بهدف تحسين الأداء على البيانات الغير متوازنة. حيث تقوم الخوارزمية بإسناد أوزان للأشجار بناء على أخطاء OOB (Out of Bag) خلال التدريب، بالتالي تحسين دقة التصنيف لكلتا الفئتين، بالإضافة لاستخدامهم الحساب المتوازي (Parallel Computing) لتقليل وقت المعالجة مع البيانات كبيرة الحجم، أظهرت هذه الخوارزمية المحسنة نتائج تتفوق على الخوارزميات الأخرى C4.5، SVM، KNN من حيث الدقة الكلية والدقة المتوازنة (Balanced Accuracy). دراسة أخرى اجراها Alejandrinoوآخرون عام (2023) [2] للمقارنة بين عدة خوارزميات تعلم خاضعة للإشراف وغير خاضعة للإشراف لتصنيف تعثر القروض باستخدام مجموعة بيانات صغيرة الحجم، شملت الخوارزميات كلاً من J48,k-NN,Bayes والانحدار اللوجستي (logistic regression)، وشملت خطوات التجهيز: تطبيع البيانات، معالجة القيم المفقودة، واستخدام SMOTE لموازنة الفئات. سجل الانحدار اللوجستي أعلى قيمة لمؤشر F1-score بمعدل 0.773. أظهرت النتائج فعالية هذه النماذج مع البيانات صغيرة الحجم ولكن تحتاج للاختبار على بيانات أكبر للتحقق من قدرتها على التعميم. وفي سياق متصل ركزت دراسة Owusu واخرون (2023) [3] على معالجة مشكلة البيانات غير المتوازنة في توقع تعثر القروض باستخدام خوارزمية ADASYN (Adaptive Synthetic Sampling) بهدف تحقيق التوازن بين الفئات، تبعها تطبيق شبكة عصبية عميقة (DNN) لتحسين الدقة التنبؤية. وقد أظهر النموذج المقترح دقة تصنيف بلغت 94.1%، متفوقًا على عدد من النماذج التقليدية الأخرى ومع ذلك، غم أن الدراسة استخدمت عددًا من المقاييس مثل Precisionو Recall و Specificityلتقييم الأداء بعد تطبيقADASYN ، إلا أنها لم تشمل مقاييس أكثر شمولًا مثل F1-score أوAUC-ROC ، والتي تُعد أكثر دقة في تقييم النماذج في ظل عدم توازن الفئات. كما أن تحليل النتائج ركّز بشكل رئيسي على مقياس Accuracy كمؤشر نهائي لتفوق النموذج، وهو ما قد يكون مضللًا في مثل هذه الحالات.

وفي هذا الإطار، قدّمت دراسة وChen واخرون حول التنبؤ بتخلف حاملي بطاقات الائتمان عن السداد باستخدام بيانات غير متوازنة في (2021) [4] نموذجا يجمع بين(Gradient Boosted Decision Tree) GBDT وتقنيات إعادة التوازن مثل
K-means SMOTE ، حيث تم تطبيق الدراسة على ثلاث مجموعات بيانات حقيقية من تايوان، ألمانيا الجنوبية، وبلجيكا. وقد تم تقييم الأداء باستخدام مجموعة شاملة من المقاييس تضمنت الدقة، Precision، Recall، F1-score، G-Mean، وAUC-ROC، حقق النموذج GBDT أعلى أداء عبر المجموعات الثلاث، حيث بلغت الدقة 88.7% لمجموعة تايوان، مما يعكس أهمية الجمع بين نماذج قوية وتقنيات موازنة فعالة، إلى جانب استخدام مقاييس تقييم متعددة لتقديم رؤية دقيقة وشاملة لأداء النموذج.

III. الداتا

مجموعة البيانات المستخدمة هي مجموعة عامة Loan Default Dataset متاحة علة منصة Kaggle [5]، تتعلق ببيانات المقترضين ومعلومات متنوعة تؤثر على قرار منح القروض، بهدف التنبؤ بالتخلف عن السداد او عدم التخلف

تتضمن البيانات سجلات تاريخية لعدد كبير من المقترضين ومعلومات مالية وديموغرافية. تتكون البيانات من 148670 سجل و34 متغير عددي وفئوي متضمنا الهدف Status وهو متغير ثنائي يأخذ القيمة (0) لعدم التخلف عن السداد و(1) في حالة التخلف عن السداد. يوجد في مجموعة البيانات عدة مشاكل كالقيم المفقودة بعدة متغيرات وبنسب متفاوتة والقيم الشاذة

outliers الموجودة في المتغيرات عددية بالإضافة الى التعدد الخطي فضلا عن عدم التوازن في المتغير الهدف.

اسم المتغير	الوصف
ID	معرف طلب القرض الخاص بالعميل
year	سنة تقديم طلب القرض
Loan limit	يحدد ما إذا كان القرض ضمن الحد النظامي (cf) أو غير نظامي(ncf)
Gender	جنس مقدم الطلب (ذكر، أنثى، مشترك، غير متوفر)
Approv in adv	يشير إلى ما إذا تم الموافقة على القرض مسبقًا pre، nopre
loan type	نوع القرض type1، type2، type3
loan purpose	الغرض من القرض p1، p2، p3، p4
Creditworthiness	الملاءة الائتمانية l1، l2
open credit	يشير إلى ما إذا كان لدى مقدم الطلب حسابات ائتمان مفتوحة opc، nopc
Business or commercial	يحدد ما إذا كان القرض لأغراض تجارية/أعمال (ob/c)أو شخصية(nob/c)
amount Loan	مبلغ القرض المطلوب
interest Rate of	معدل الفائدة المفروضة على القرض
Interest rate spread	الفرق بين معدل الفائدة ومعدل معياري (مؤشر قياسي)
Upfront charges	الرسوم المسبقة المرتبطة بالحصول على القرض
term	مدة القرض بالأشهر
Neg amortization	يشير إلى ما إذا كان القرض يسمح بالسداد السلبي not_neg، neg_amm
interest only	يشير إلى ما إذا كان القرض يحتوي على خيار دفع الفائدة فقط int_only، not_int
Lump sum payment	يشير إلى ما إذا كان يُطلب دفع مبلغ مقطوع في نهاية مدة القرض lpsm، not_lpsm
property value	قيمة العقار الممول بالقرض
construction type	نوع البناء - sb بناء في الموقع، - mhمنزل مُصنّع
occupancy type	نوع الإشغال - pr إقامة أساسية، - srإقامة ثانوية، ir -عقار استثماري
Secured by	نوع الضمان المستخدم لتأمين القرض - home منزل، - landأرض
total units	عدد الوحدات في العقار الممول1U) ، U2، U3، U)4
income	الدخل السنوي لمقدم الطلب
credit type	نوع تقرير الائتمان لمقدم الطلب CIB ، CRIF، EQUI ،EXP
Credit Score	الدرجة الائتمانية لمقدم الطلب
co-applicant credit type	نوع تقرير الائتمان للمشارك في القرض CIB، EXP
age	عمر مقدم الطلب
Submission of application	طريقة تقديم الطلب - to_inst إلى المؤسسة، - not_instليس إلى المؤسسة
LTV	نسبة القرض إلى قيمة العقار(Loan to Value)
Region	المنطقة الجغرافية للعقار (الشمال، الجنوب، الوسط، الشمال الشرقي)
Security Type	نوع الضمان أو الأصل الذي يؤمن القرض (مباشر، غير مباشر)
Status	الحالة: هل تم التخلف عن السداد؟ (1: تخلف، 0: سداد)
dtir1	نسبة الدين إلى الدخل(Debt-to-Income Ratio)

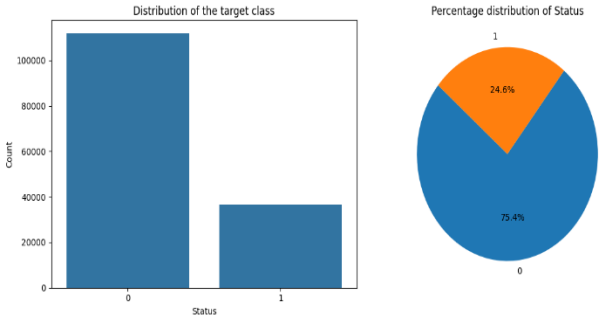
الجدول 1: وصف المتغيرات في مجموعة البيانات

IV. الاستكشاف والتحليل

قبل البدء بمعالجة مجموعة البيانات قمنا بإجراء تحليل إحصائي بهدف فهم البنية العامة وتوزيع المتغيرات وفهم الأنماط.

i. توازن مجموعة البيانات

بداية فحصنا توزع مجموعة البيانات أظهر توزع الداتا نسبة غير متوازنة في المتغير الهدف كما يظهر في (الشكل 1)، حيث تمثل نسبة حالات التخلف عن السداد ما يقارب 24.6% فقط وهذا ما يستدعي تطبيق تقنيات إعادة التوازن أثناء بناء النماذج التنبؤية لاحقًا.



الشكل 1: توزع مجموعة البيانات

ii. القيم المفقودة

تتضمن مجموعة البيانات قيما مفقودة في كل من المتغيرات الرقمية والفئوية على حد سواء ويتفاوت عددها بين متغير واخر حيث يظهر جدول القيم المفقودة (الجدول رقم 2) ان اقلها عددا بلغ عددها 41 في متغير term واكبرها عددها يساوي 39642 لدى upfront changes ولتحديد كيفية التعامل مع هذه القيم يتطلب ذلك فحص ان كانت هذه القيم مفقودة بشكل عشوائي MCAR ام انها غير عشوائي NMAR , ولاكتشاف ذلك يجب ان نحدد المتغير الذي يجب ان نقارن معه لتحديد علاقتها , من الجدير بالذكر ان العدد الأكبر من القيم المفقودة البالغ عددها 39642 ومتغيرات أخرى تحوي قيم متشابهة قريبة جدا من عدد العينات في فئة التخلف عن السداد في المتغير الهدف , لذا وجدنا انه من الضروري فحص ان كان الفقد عشوائي او غير عشوائي اعتمادا على الهدف و بعد عرض القيم المفقودة في كل متغير بالنسبة للمتغير الهدف وجدنا ان جميع المتغيرات الفئوية غير مرتبطة به وفقدانها يتم بشكل عشوائي.

بينما المتغيرات الرقمية ففقدان البيانات غير عشوائي ومرتبط بالهدف ولكن بنسب متفاوتة.

كل من "interest of spread", "rate Interest rate" تكون مفقودة دائما عندما يكون الحالة تخلف عن السداد وتمتلك قيم فقط عند السداد بينما "charges Upfront" لديه حالة مشابه ولكن معظم القيم المفقودة في حالة عدم السداد وجزء صغير من السجلات يحوي قيم فقط وايضا يوجد عدد من حالات السداد التي تختفي فيها قيمة المتغير.

وأخيرًا، يوجد قسم من المتغيرات التي تحتوي على قيم مفقودة في كلا الصنفين، لكن أحد الصنفين يحتوي على عدد أكبر من القيم المفقودة مقارنة بالصنف الآخر مثل "income", "property value","dtirl","LTV".

المتغير	عدد القيم	نوعه
Upfront charges	39642	عددي
Interest rate spread	36639	عددي
Rate of interest	36439	عددي
dtir1	24121	عددي
Property value	15098	عددي
LTV	15098	عددي
income	9150	عددي
Loan limit	3344	فئوي
Approv in adv	908	فئوي
age	200	فئوي
Submission of application	200	فئوي
Loan purpose	134	عددي
Neg_ammortization	121	فئوي
term	41	عددي

الجدول 2: جدول القيم المفقودة

iii. القيم الشاذة

المتغيرات العددية أظهرت انحرافات ملحوظة وتوزيعات غير طبيعية بالإضافة للتباين الكبير في القيم.

حيث كل من loan amount, Upfront charges, income, property) (value لديها توزع غير طبيعي وانحراف يميني كبير.

وكل من (Income, LTV) تمتلك قيم شاذة كبيرة تشوه التوزيع، ففي income هناك دخل بقيمة 0، وقيمة قصوى ضخمة (578,580) مما يدل على احتمالية وجود دخل مفقود أو مدخل بشكل خاطئ، اما LTVلقيمة

القصوى 7831.25 وهذه تشير إلى قيمة غير صالحة وغير منطقية. أيضا المتغير termيفتقد إلى التنوع، إذ تتكرر قيمة 360شهرًا بنسبة تتجاوز% 75.

iv. المتغيرات الفئوية

معظمها يكون بين فئتين الى أربع فئات ومتغير واحد فقط يحوي على 7. بعض المتغيرات الثنائية أحد فئاتها مهيمنة على الأخرى مثل (occupancy type, security type، open credit، secured by) حيث الفئة النادرة فيها اقل من 1% من القيم تتطلب هذه المتغيرات دراسة تأثيرها على النماذج وتجربة حذفها او تشفيرها بطريقة مناسبة. أيضا متغير total uniteيملك 4 فئات وفئة واحدة فقط مهيمنة يمكن محاولة دمجها معا ورؤية تأثيرها.

v. العلاقات بين المتغيرات

الارتباطات الإحصائية:

لم تظهر أي من المتغيرات العددية ارتباطًا قويًا بالمتغير الهدف، حيث جميع معاملات ارتباط هذه المتغيرات بالمتغير الهدف ضعيفة.

المتغيرات term, ID, yearغير دالة إحصائيًا.

أجرينا اختبار كاي-تربيع على المتغيرات الفئوية والهدف وجدنا ان المتغيرات جميعها يملك دلالة احصائيا ولكن (occupancy type, security type، open credit، secured by) ذات دلالة إحصائية ضعيفة مع الهدف.

اما (Neg ammortization, Lump sum payment) ومتغيرات أخرى كانت دلالتها الاحصائية عالية.

تفاعل المتغيرات وتحليل التأثيرات المشتركة:

عند تحليل العلاقات الثنائية والثلاثية بين المتغيرات، لوحظت بعض الأنماط المهمة التي تساعد في تفسير ديناميكية البيانات، وتُشير إلى ارتباطات ذات دلالة بين بعض الخصائص.

العلاقة بين loan amount وrate of interest وبين property valueو LTV علاقة عكسية ضعيفة بالتالي لا يوجد تعدد خطي قوي بين المتغيرين.

بالمقابل هناك علاقة طردية قوية بين property value وloan amount ما يدل على وجود تعدد حطي بينهما.

أظهرت المنطقة الجنوبية نشاطًا ملحوظًا في طلبات القروض، وكانت الفئة (type1)من loan type هي الأكثر شيوعًا، في حين سُجّلت أدنى نسبة نشاط في المنطقة الشمالية الشرقية

توزيع قيم (loan amount)كان متقاربًا بين فئات loan type الثلاثة من حيث الوسيط والنطاق البيني(IQR) ، بينما ظهر وجود قيم شاذة عالية بشكل أكبر في النوعين type3,type1 .

بالنسبة ل rate of interest فقد لوحظ اختلاف واضح بين القروض التي تتضمن تسديدًا سلبيً (Neg Amortization) وتلك التي لا تتضمنه، حيث كانت معدلات الفائدة أعلى وأكثر تباينًا في القروض ذات التسديد السلبي، مما يشير إلى مخاطرة وتكلفة مالية أكبر.

أيضا القروض ذات التسديد السلبي تميل إلى أن تكون بقيمة أقل مقارنة بالقروض العادية، وكلا النوعين يتمركز في مبالغ صغيرة إلى متوسطة.

معظم القروض تُمنح لأشخاص في الفئة العمرية 44–25عامًا

كذلك، العلاقة بين الدخل income وقيمة القرض loan amountكانت إيجابية لكنها ضعيفة، حيث يميل أصحاب الدخل المرتفع إلى الحصول على قروض أكبر. الأشخاص الذين سدّدوا قروضهم كانوا يمتلكون معدلات income، وcredit score، ونسبة LTV أعلى، مقارنة بمن تخلّفوا عن السداد.

عند سداد القرض يكون income وcredit score وloan amount وLTV مرتفعة.

v. المعالجة المسبقة وهندسة السمات

i. معالجة القيم الشاذة والتوزع

المتغيرات اتي تملك توزع غير طبيعي وقيم موجبة (loan amount, Upfront charges, income, property value)استخدمنا التحويل اللوغارتمي لجعل هذه المتغيرات تتبع التوزع الطبيعي للبيانات .

اما القيم الشاذة للتقليل من تأثيرها دون حذفها لأنه البيانات غير متوازنة وحذفها يمكن ان يزيد من عدم التوازن.

م اعتماد تقنية القص(Clipping) ، وهي طريقة تعتمد على تحديد حدود دنيا **وعليا** استنادًا إلى القيم المئوية(quantiles) ، ثم يتم استبدال القيم التي تتجاوز هذه الحدود بالقيم الحدية نفسه .

هذه المعالجة ساهمت في تقليل التشتت والانحراف داخل التوزيع، مع الحفاظ على حجم البيانات الكامل دون حذف السجلات

ii. التعامل مع القيم المفقودة

القيم الفقودة في الداتا نوعين متغيرات عددية غير عشوائية مرتبطة بالهدف ومتغيرات فئوية عشوائية .

بالنسبة للمتغيرات الفئوية فقد تم التعامل مع هذه القيم باستخدام القيم الأكثر تكرارا حيث ان نسبة هذه القيم لم تكن كبيرة .

اما المتغيرات العددية حيث القيم المفقودة تعتمد على متغير الهدف، فإن هذا النوع من الفقدان يكون "معلوماتيًا" ويُحتمل أن يحمل إشارات قوية عن النتيجة. ولكن استخدام الهدف في نماذج الإتمام (imputation) خلال النشر أو الاستخدام العملي غير ممكن (لأن الهدف غير متاح في وقت التنبؤ).

في هذا السياق قدم Sisk واخرون[7] طرق للتعامل مع البيانات المفقودة وكانت الطريقة المستخدمة مع الفقدان الغير عشوائي المعتمد على الهدف هي استخدام الإتمام بالانحدار Regression Imputation وتجاهل الهدف أي الهدف لا يدخل بعملية ملئ القيم وأيضا إدراج مؤشر للفقدان (missing indicator) حيث تستخدم هذه الطريقة عندم يسمح بوجود قيم مفقودة عند التشغيل وقد طبقنا هذه الطريقة على المتغيرات العددية لدينا ولكن بالنسبة للمتغيرات "rate Interest rate", "interest of spread", "Upfront charges" فقد حذفنا هذه المتغيرات لأنه وجودها يعد مؤشر للهدف حيث الفئة الأقل بالهدف جميع قيمها فارغة فحتى مع استخدام الأتمام بالانحدار او المؤشرات او حتى تركها دون معالجة او ملءها بقيم ثابتة فانها تؤدي الى دقة 100% في النماذج بالتالي تم حذفها.

iii. التشفير

لتحويل المتغيرات الفئوية إلى شكل رقمي استخدمنا ORDENAL ENCODER لكل من المتغيرات الثنايئة والمتغيرات التي تملك فئات بترتيب طبيعي.

كل من عمليات معالجة القيم المفقودة والتشفير تم تضمينها في PIPLINE أيضا بالنيبة للمتغير TOTAL UNITE قمنا بدمج الفئات النادرة مع بعضها لتقليل التشتت وزيادة عدد العينات ضمن كل فئة بعد الدمج، مما يُحسن من استقرار النموذج.

iv. هندسة السمات

تضمنت مرحلة هندسة السمات إجراءين رئيسيين **:**إزالة السمات غير المفيدة، **و**إنشاء سمات جديدة مشتقة يمكن أن تسهم في تحسين أداء النماذج.

في البداية، تم حذف بعض المتغيرات التي ثبت ضعف فائدتها للنمذجة، إما لأنها معرفات أو ذات تنوع منخفض جدًا، أو أنها لم تظهر كسمات مهمة خلال تحليل الأهمية، وهي:

ID, year: تمثل معرفًا وزمن التسجيل لقيمة واحدة ، ولا تحمل دلالة تنبؤية مباشرة.

term : رغم ارتباطها بطول القرض، إلا أن أكثر من 70% من القيم كانت 360، مما قلل من تنوعها، كما أن الاختبارات الإحصائية بيّنت ضعف علاقتها بالهدف.

Property value :رغم أهميته النظرية، إلا أن النماذج التي تم تدريبها بوجود هذا المتغير أظهرت مؤشرات واضحة لفرط التكيّف (Overfitting)، لذلك تم حذفه.

كما تم حذف أربع متغيرات فئوية ثنائية هي:

Security Type, secured by, open credit, construction type، وذلك بسبب هيمنة إحدى الفئات وندرة الأخرى (أقل من 1%) ، مما يسبب خللاً في التوزيع يصعب على النماذج تعلمه بفعالية. كما أن اختبار كاي-تربيع أظهر دلالة إحصائية ضعيفة لهذه المتغيرات مع الهدف.

من جهة أخرى، تم اشتقاق عدد من المتغيرات الجديدة بهدف تمثيل العلاقات المعقدة داخل البيانات بشكل أكثر فائدة للنماذج:

loan per age **:**يمثل نسبة مبلغ القرض إلى عمر المقترض، مما يساعد على فهم مدى العبء المالي بالنسبة للفئة العمرية، خاصة أن المقترضين الأصغر سنًا قد يمثلون مخاطرة أعلى عند طلب قروض كبيرة.

LTV High flag: متغير ثنائي يشير إلى تجاوز نسبة القرض لقيمة العقار حد 0.85، وهو مؤشر تقليدي للمخاطرة في مجال القروض العقارية.

risk score **:** مؤشر رقمي تم بناؤه من دمج ثلاث خصائص عالية الخطورة هي: التسديد السلبي(Neg_ammortization) ، السداد الجزئي (interest only) ، والدفع الإجمالي .(lump sum payment) تم ترميز كل متغير وتحويله إلى مقياس عددي ثم جمعها لتكوين درجة مركبة للمخاطرة.
هذه السمات المشتقة ساعدت على تمثيل الأنماط المالية والسلوكية للمقترضين بطريقة مبسطة وفعالة، وساهمت في تحسين أداء النماذج وزيادة قدرتها التنبؤية.

VI. النتائج والتقيم

في هذا القسم، نستعرض نتائج النماذج التنبؤية المستخدمة، مع التركيز على المقارنة بينهم من حيث الدقة والأداء والتعقيد.
قمنا لتقسيم مجموعة البيانات الى 3 اقسام تدريب وتحقق واختبار، مع الحفاظ على نفس توزيع الفئات باستخدام Stratified Split.
ولتحقيق أداء أفضل، تم استخدام مكتبة Optuna [7]لضبط معلمات النماذج تلقائيًا باستخدام تقنية التحقق المتقاطع.(Cross-Validation) .
اما طرق التقييم اعتمدنا على مجموعة من المقاييس شملت:

i. الدقة (Precision) :لقياس نسبة التوقعات الصحيحة من الإيجابيات المتوقعة.
ii. الاسترجاع (Recall) **:**لقياس القدرة على استرجاع الحالات الفعلية.
iii. معدل F1 (F1-score) :المتوسط التوافقي بين الدقة والاسترجاع.
iv. PR AUC :المساحة تحت منحنىPrecision-Recall ، وهي أكثر دقة في حالات عدم توازن البيانات.
v. مصفوفة التعارض (Confusion Matrix) :لتحليل الأخطاء في كل فئة.

وقمنا بتجربة عدة استراتيجيات لمعالجة عدم توازن الفئات في المتغير الهدف، شملت كل من:
ADASYN, Class Weighting, SMOTE + Tomek , SMOTE
تم تطبيق هذه الطرق على نموذجين مختلفين (Decision Tree وXGBoost) ، ويُبين الجدولان (3) و(4) نتائج أداء كل نموذج مع كل تقنية موازنة.

إعادة التوازن	precision	recall	F1 score	PR AUC
SMOT	**0.87**	**0.58**	**0.70**	**0.78**
ADASYN	0.87	0.57	0.69	0.77
Class weight	0.87	0.57	0.69	0.77

الجدول 3: أداء نموذج Decision Tree تحت استراتيجيات إعادة التوازن

إعادة التوازن	precision	recall	F1 score	PR AUC
SMOT	**0.91**	**0.61**	**0.7346**	**0.83**
ADASYN	0.91	0.61	0.7326	0.83
Class weight	0.75	0.70	0.72	0.83
SMOTE+Tomek	0.91	0.61	0.7341	0.83

الجدول 4: أداء XGBoostتحت استراتيجيات إعادة التوازن

يتضح من النتائج أن نموذج XGBoostيتفوق بشكل ملحوظ على نموذج أشجار القرار من حيث جميع مقاييس الأداء، خاصة في Precisionو-F1 score. كما أن جميع طرق إعادة التوازن حسّنت الأداء، إلا أن SMOTEو SMOTE+Tomekأظهرا نتائج متميزة عند استخدام XGBoost. من جهة أخرى، حافظت أشجار القرار على أداء مستقر نسبيًا بغض النظر عن طريقة إعادة التوازن، ولكنها تبقى أقل فعالية من النماذج الأكثر تقدمًا.

قمنا بتجربة عدة نماذج ومقارنة اداءها وذلك باستخدام SMOT لموازنة الفئات

النموذج	precision	recall	F1 score	PR AUC
Decision Tree	0.87	0.58	0.70	0.78
XGB	0.91	0.6133	0.7346	0.8326
Catboost	0.92	0.6100	0.7353	0.8329
lightGBM	0.90	0.6189	0.7376	0.8316
adaBoost	0.89	0.5111	0.6507	0.7775
Voting (Xgb+lgbm+catb)	0.92	0.6153	0.7376	0.8352
Stacking (Xgb+lgbm+catb +dt)	**0.8718**	**0.6508**	**0.7453**	**0.8348**
Random forest	0.92	0.58	0.71	0.8127
Random forest+oob	0.90	0.59	0.72	0.80
bagging	0.9117	0.5899	0.7163	0.8180
Extra trees	0.88	0.5558	0.6842	0.7925

الجدول 5: مقارنة اداء النماذج

تمت تجربة عدة تكوينات لنموذج Stackingوالتي تعتمد على دمج تنبؤات عدة نماذج أساسية:

- التكوين الأول.XGBoost + LightGBM + CatBoost
- التكوين الثاني: الثلاثة أعلاه مع إضافة Decision Tree.
- التكوين الثالث CatBoost + LightGBM + Decision Tree.

وقد أظهرت التجربة الثانية أفضل توازن بين F1-score والاسترجاع، مما يشير إلى أن وجود شجرة قرار بسيطة كمكمّل للنماذج المعقدة قد يعزز التعميم.

أيضا قمنا بتحسين نموذج Random Forestمن خلال تخصيص أوزان لكل شجرة باستخدام أخطاء العينات خارج الحقيبة Out-of-Bag) (Errors، وهي الطريقة المقترحة في [1] لتحسين دقة التصنيف. تم منح وزن أعلى للأشجار ذات الأداء الأفضل في الـOOB ، ثم إجراء تصويت مرجّح. أظهرت هذه الطريقة تحسنًا طفيفًا على أداء النموذج الأساسي.

مقارنة النتائج من (الجدول 5):

- أفضل F1-scoreسُجل في نموذج **+** Stacking (XGB LGBM + CAT + DT)يليه مباشرة نموذج .LightGBM
- نموذج CatBoostوXGBoost قدما أداء متقاربًا جدًا.
- نموذج Baggingأظهر دقة جيدة لكن تراجعت مقاييس الاسترجاع.
- Decision Treeو AdaBoostسجلوا أداء أقل من النماذج الأخرى.

النموذج	التعقيد	الزمن
Decision Tree	منخفض	4ثانية
XGB	متوسط	2 ثانية
Catboost	متوسط	2 ثانية
lightGBM	متوسط	2 ثانية
adaBoost	متوسط	2 ثانية

Voting (Xgb+lgbm+catb)	مرتفع	1 دقيقة
Stacking (Xgb+lgbm+catb+dt)	مرتفع	2دقيقة
Random forest	مرتفع	1 دقيقة
Random forest+oob	مرتفع	1 دقيقة
bagging	متوسط	2 ثانية
Extra trees	متوسط	2 ثانية

الجدول 6: مقارنة تعقيد النماذج

- النماذج المجمعة مثل Stackingتتطلب وقتًا أطول بسبب تدريب عدة نماذج فرعية.
- LightGBMقدّم توازنًا ممتازًا بين الدقة والسرعة.
- Stackingاستفاد من تجميع نقاط القوة في عدة نماذج، مما عزز التعميم والتوازن.
- LightGBMأظهر تفوقًا ملحوظًا بفضل كفاءته العالية مع البيانات الكثيرة والمتغيرات المتنوعة، واستغلاله الجيد للبنية الهرمية للبيانات.
- Decision Treeبقي الخيار الأبسط والأسرع، مما يجعله مناسبًا كنموذج خط أساس (baseline) للمقارنة، لكنه افتقر للدقة في الحالات المعقدة.
- Random Forest OOB Votingحسّن النموذج عبر استبعاد الأشجار الضعيفة، لكنه لم يتفوق على النماذج المعقدة المجمعة.

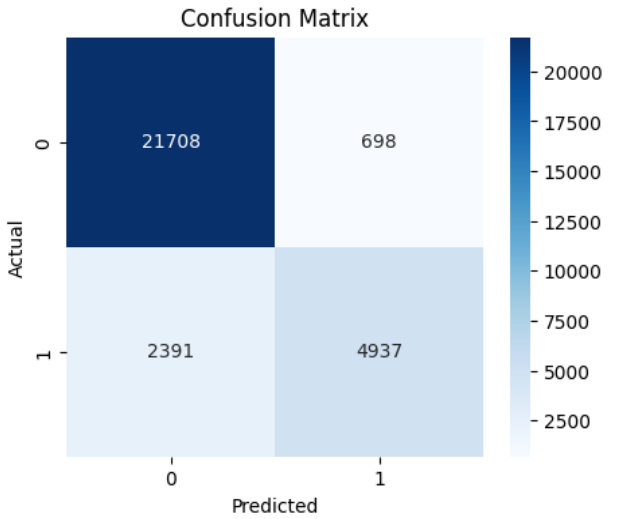
أداء النموذج على بيانات الاختبار:

بعد الانتهاء من تدريب النماذج وضبط المعلمات باستخدام بيانات التحقق (Validation)، تم تقييم أفضل نموذج Stackingعلى مجموعة بيانات الاختبار.

حقق نموذجStacking على بيانات الاختبار النتائج التالية (الجدول رقم 7) :

Precision	0.8761
Recall	0.6737
F1 score	0.7637
PR AUC score	0.8463
Accuracy score	0.8961
Balanced accuracy score	0.8213

الجدول 7: نتائج نموذج stackingعلى بيانات الاختبار



الشكل 2: مصفوفة التعارض لنموذج stacking

المقارنة مع الاعمال السابقة :

النموذج	F1 score	Balanced accuracy	accuracy	recall
[1] RF+OOB	-	0.765118	0.864929	-
[2] lBk (k-NN)	0.780	-	-	-
[3] adasyn+dnn	-	-	94.1	0.972
[4] random forest	0.57491	-	0.65550	0.57830
Stacking (نحن)	0.7637	0.8213	0.8961	0.6737

الجدول 8: مقارنة النتائج مع الاعمال السابقة

بعد المقارنة بين نموزجنا والاعمال السابقة وجدنا ان

بناءً على التحليل المقارن بين عملنا والدراسات السابقة، نستخلص ما يلي:

أظهر نموذج الـ Stackingالمستخدم في دراستنا أداءً متوازنًا من حيث F1-score (0.7637)و Balanced Accuracy (0.8213) و(Accuracy (0.8961، متفوقًا على بعض الدراسات مثل [4] التي استخدمت Random Forest وحققت F1-score منخفضًا بلغ 0.57491 .

رغم أن الدراسة [3] التي استخدمت ADASYN مع شبكة عصبية عميقة (DNN)أبلغت عن Accuracy بلغت 94.1% و Recallبنسبة 97.2%، فإن غياب مؤشرات مثل F1-score و Balanced Accuracyيثير مخاوف حول التحيز الناتج عن عدم التوازن في البيانات. في المقابل، نموذجنا قدم نتائج متوازنة دون التضحية بعدد كبير من الأخطاء من الفئة النادرة .

على الرغم من أن بعض الدراسات (مثل [3] و[4]) استخدمت مجموعات بيانات أكبر من مجموعتنا، إلا أن أداء نموذجنا كان منافسًا أو متفوقًا، مما يشير إلى فعالية استراتيجية الدمج التنبؤي (Stacking) في تحقيق تعميم جيد حتى على مجموعات بيانات متوسطة الحجم .

استخدام تقنيات مثل SMOTEفي مرحلة التدريب ساعد في معالجة عدم التوازن في المتغير الهدف، مما ساهم في تحسين أداء النموذج على الفئة الأقل تمثيلًا (القروض المتعثرة).

فرص التحسين المستقبلية

تحسين قيمة الـ Recall (0.6737) للكشف عن مزيد من حالات التعثر . وتعزيز خطوات هندسة الخصائص لرفع دقة التنبؤ. وأيضا تجربة خوارزميات أخرى كالشبكات العصبية.

VII. خاتمة

في هذه الدراسة، تم تطوير إطار عمل تنبؤي متكامل لتصنيف حالات تعثر سداد القروض باستخدام خوارزميات تعلم الآلة. شملت منهجية العمل معالجة متقدمة للبيانات، بما في ذلك التعامل مع القيم المفقودة والشاذة، بالإضافة إلى تصميم سمات مشتقة تساهم في تعزيز القدرة التنبؤية للنماذج. أظهرت نتائج التجريب أن نماذج مثل LightGBM و Stackingتحقق توازنًا ممتازًا بين الأداء والدقة والاسترجاع، مع وقت تدريب مقبول. كما ساهم استخدام Optuna في تحسين الضبط التلقائي للنماذج، وأدى إلى رفع دقة التنبؤ بشكل ملحوظ.

مقارنة بالدراسات السابقة، تُظهر النماذج المقترحة أداء تنافسيًا أو متفوقًا في بعض المؤشرات، مما يدل على فعالية الأساليب المستخدمة في هذه الورقة.

مستقبلًا، يمكن توسيع هذا العمل ليشمل نماذج أكثر تعقيدًا مثل الشبكات العصبية التفسيرية، أو دمج مصادر بيانات خارجية مثل سجل الائتمان أو سلوك الإنفاق.

VIII. المراجع

[1] H. Zhou and Z. Wang, "Loan default prediction on large imbalanced data using random forests," *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 10, no. 6, pp. 1519–1525, Oct. 2012, doi: 10.11591/telkomnika. v10i6.1323.

[2] J. C. Alejandrino, J. P. Bolacoy, and J. V. Murcia, "Supervised and unsupervised data mining approaches in loan default prediction," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 1837–1847, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1837-1847.

[3] E. Owusu, R. Quainoo, S. Mensah, J. K. Appati, " A Deep Learning Approach for Loan Default Prediction Using Imbalanced Dataset," International *Journal of Intelligent Information Technologies (IJIIIT)*, 19(1), 1–15, 2023, doi: 10.4018/IJIIT.318672

[4] Y.-R. Chen, J.-S. Leu, S.-A. Huang, J.-T. Wang, and J. Takada, "Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets," *IEEE Access*, vol. 9, pp. 73108–73117, May 2021. doi: 10.1109/ACCESS.2021.3079701

[5] M. Yasser H., *"*Loan Default Dataset*," Kaggle*. [Online]. Available: https://www.kaggle.com/datasets/yasserh/loan-default-dataset. [Accessed: Jul. 5, 2025]

[6] R. Sisk, M. Sperrin, N. Peek, M. van Smeden and G. P. Martin, "Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study," *Statistical Methods in Medical Research*, vol. 32, no. 8, pp. 1461–1477, 2023. doi:10.1177/09622802231165001

[7] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD ’19)*, Anchorage, AK, USA, Aug. 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.