التحليل الإحصائي لبيانات طالبي القروض لتقييم جدارتهم الائتمانية

Statistical analysis to the data of consumer credit to evaluate credit scoring

د. رزق السيد حامد الوزير*

ملخص

يُستَخدم مصطلح تقييم الائتمان (الجدارة الائتمانية) في صناعة الائتمان لوصف الأساليب المستخدمة لتصنيف طالبي القروض الجدد في فئتي خطر وفقا لسلوكهم المحتمل في السداد: "متعثر" في السداد defaultedو"غير متعثر"non-default. ويحقق التصنيف الجيد منفعة مزدوجة لكل من الدائن (زيادة ربحه أوتقليص خسارته) وطالب القرض (تجنب التعثر).ويهدف البحث لتقييم قاعدة بيانات القروض الشخصية لنادي الإقراض Lending Club، والتنبؤ بما إذا كان مقدم طلب الائتمان الجديد سيسدد قرضه (جدير بالائتمان creditworthy) أم سيتعثر فيه (غير جدير بالائتمان non-creditworthy)باختبار البيانات التاريخية لقروض المستهلكين لعينة حجمها 42539 حالة استلموا قروضهم خلال الفترة ٢٠١١-٢٠٠٧ باستخدام كل من نماذج الانحدار اللوجستي وشجرة القرارات، وتقديم نموذج اتوماتيكي يُستند عليه في تقييم الائتمان. وتتكون مجموعة البيانات من الخصائص المالية وخليط من القروض المسددة والمتعثرة للعملاء الجديرين بالائتمانمن وجهة نظرالنادي. كما شملت مجموعة البيانات أيضاً الخصائص المالية للعملاء غير الجديرين بالائتمان (والذين تُرفض طلباتهم في العادة ولا تُسجل بياناتهم في غالبية قواعد البيانات)، وهو ما يقضي على التحيز الذي تعاني منه معظم دراسات تقييم الائتمان بسبب اقتصارها على دراسة بيانات القروض المقبولة فقط. ويقارن البحث معدلات دقة التصنيف الصحيح للانحدار اللوجستي وثلاث خوارزميات لشجرة القرارات. وقد تبين تفوق الانحدار اللوجستي على طرق أشجار القرار في التصنيف التنبؤي لمجموعة البيانات. وكانت أهم المتغيرات المؤثرة على حالة القرض: الدفعات الكلية، ومبلغ القرض، ودرجة التصنيف الائتماني، وأجل القرض، ومطابقة خصائص العميل للسياسة الائتمانية للبنك، والغرض من القرض، وحد التقاطع على الترتيب.

الكلمات الدالة: الجدارة الائتمانية، الانحدار اللوجستي، شجرة التصنيفات.

Summary:

Credit scoring (creditworthy) is the term used by the credit industry to describe methods used for classifying applicants for credit into two risk classes according to their likely repayment behavior: "default" and "non-default".Accurate classification is of benefit both to the creditor (in terms of increased profit or reduced loss) and to the loan applicant (avoiding over-commitment).The research aims to evaluate theLending Club database for consumer credit by examining the historical data of consumer loans of size 42539during the period 2007-2011 using logistic regression and decision trees and to predict whether a new applicant paid off or defaulted upon his/her loan using automatic model.The data set consists of the financial attributes of each customer and includes a mixture of loans that the customers paid off or defaulted upon for both accepted and rejected loans, which eliminates the bias suffered by the most credit assessment studies due to limiting themselves to the study of loan data only accepted. The paper compares the correct classification accuracy rates for logistic regression andthree decision tree algorithms. The logestic regression outperformed the decision tree methods. The most important variables were: total payments, amount, grade, term, policy, purpose and intercept.

***Keywords:** Credit Scoring, Logistic Regression, Decision Trees.*

*أستاذ مساعد الإحصاء التطبيقي، قسم اقتصاديات وإدارة المشروعات،كلية العلوم الإدارية والمالية، جامعة الطائف–مدرس الإحصاء التطبيقي، قسم الإحصاء التطبيقي والتأمين، كلية التجارة، جامعة المنصورة.

١) مقدمة البحث

١-١ مقدمة

ارتفع معدل نمو الدخل الفردي في السعودية في العشر سنوات الأخيرة بشكل ملحوظ، وقد صاحب ذلك زيادة كبيرة أيضاً في حجم المدخرات وهو ما شجع البنوك (الأجنبية والوطنية) على دخول السوق السعودية لتستفيد من هذا الوفر الذي لا تدفع عليه في الغالب أي فوائد للمدخرين. ومن ثم، فقد تسابقت البنوك في تقديم برامج تسويقية مختلفة ساهمت –إلى جانب ارتفاع نفقة المعيشة والنمو الاقتصادي الاجتماعي– في زيادة نسبة طالبي القروض.فقد ارتفعت القروض الشخصية من ١٧٨ مليار ريال في ٢٠٠٧ إلى ٢٨٥,٤ مليار في الربع الثالث من عام ٢٠١٢ (مؤسسةالنقد العربي السعودي، التقرير السنوي، ١٩ فبراير ٢٠١٣) بزيادة ٦٠% في ٥ سنوات، وهو ما جعل السعودية تحتل المركز الأول على مستوى الخليج بالنسبة للقروض الشخصية.

وعلى الجانب الآخر، فقد واكب هذا النمو الاقتصادي تطور صناعة الخدمات المالية أيضاً بشكل سريع.إذ تنافست البنوك في تقديم عروض مغرية على قروض المستهلك بمعدل فائدة يتراوح بين ١% إلى ٣% (حسب السجل الائتماني للعميل) وبفترة سداد تصل إلى ٥ سنوات. وتشترط البنوك (لمنح الائتمان في السعودية): تحويل الراتب، والتاريخ الائتماني الجيد للعميل، وأن تكون جهة العمل ضمن الجهات الموقعة على التعامل مع البنك. وكانت أغراض خفض معدل الفائدة على الإقراض: تحريك السيولة المكدسة لدى البنوك التي لا تغرم عليها أي تكاليف، ومحاولة الفوز بأكبر عدد ممكن من العملاء وخاصةً بعد انتهاء سداد القروض المعروفة بقروض الأسهم، وتحويل الرواتب للبنك.

ولا تقتصر الجدارة الائتمانية على الأفراد، بل أنها مصطلح يشمل أيضاً الدول ومنشآت الأعمال، غير أن اختلاف المتغيرات الحاكمة لكل من الحالات الثلاثة جعلها تُعالج في البحوث التطبيقية كل على حده. ويدرس البحث الحالة الأولى المعروفة باسم القروض الشخصية (قروض المستهلك/قروض الأفراد/قروض التجزئة).

١-٢ مشكلة البحث

في ضوء العرض المتزايد لكمية النقود في البنوك ومع الانخفاض الملحوظ لسعر الفائدة على الإقراض في السنوات الأخيرة وارتفاع نفقة المعيشة، يتقدم آلاف الأفراد يومياً للبنوك بطلبات للحصول على قروض جديدة. فما يلبس أن ينتهي العميل من سداد قرضه القديم (وفي بعض الأحيان نصفه)، إلا ويتقدم بطلب للحصول على قرض جديد. ومع تحقيق البنوك لأرباح من وراء هذه العملية، إلا أن خسائرها أيضاً كانت جسيمة بسبب عدم قدرة بعض المقترضين على السداد نتيجة التقييم الخاطئ للائتمان. فقد ارتفعت خسائر البنوك السعودية بسبب القروض الشخصية من ٥% (١٠ مليار من ٢٠٢ مليار ريال) في عام ٢٠٠٨ إلى ١٦% (٤٧ مليار من ٢٨٥,٤ مليار) عام ٢٠١٢. وتحتاج البنوك إلى نظام لتلبية تلك الطلبات أو رفضها. وعلى الرغم من أن غالبية البنوك العالمية تستخدم الآن الأنظمة الأوتوماتيكية لتقييم الائتمان استجابةً للقواعد التي أقرتها لجنة باسل للبنوك المركزية، إلا أن هذا التقييم في بعض البلدان العربية مازال يعتمد على الأحكام والمعرفة الشخصية والطرق الإحصائية

التقليدية.واستجابة لذلك، يجب على الجهات المانحة للائتمان استخدام وتطوير نماذج تقييم الائتمان إلى جانب الطرق الإحصائية التقليدية لدعم قراراتها الائتمانية بهدف زيادة الدقة المتعلقة بقرارات منح القروض، بحيث يُمنح الائتمان للمتقدمين الأكثر جدارة (وهو ما يؤدي لزيادة الربح)، ويُحجب عن المتقدمين الأقل جدارة (وهوما يؤديلتقليص الخسارة).

ويُعد تقييم قرارات خطر الائتمان عملية معقدة بطبيعتها بسبب العلاقات غير الخطية بين المتغيرات المستقلة التي تتفاعل مع بعضها البعض إلى جانب أشكال الخطر المختلفة التي تتضمنها تلك العملية. وعلى الرغم من أن هذا التقييم قد يتأثر ببعض العوامل التي يصعب قياسها (كالزواج المفاجئ أو الإصابة)، إلا أن بعض القروض السيئة يمكن تجنبها باستخدام أساليب تقييم يتوفرلها قدره أكثر على تمييز حالتي الخطر وهو أمرمرغوب للغاية.

١-٣ أهمية البحث

يُتوقع أن يساهم البحث في تقليل نسبة التعثر في سداد القروض لأنه يتناول نماذج تقييم الائتمان وفق أحدث طرق التنقيب في البيانات–التي لم تطبقها معظم البنوك بعد. ولا شك أن ذلك سيعود بالنفع –في حالة إقرار تلك الطرق– على كلٍ من:

–المقترض:بتوضيح أسباب رفض قرضه بدلا من الموافقة عليه على الرغم من عدم جدارته، وتعثره، واستبدال القرض القديم بآخر جديد وغرقه في الديون. كما يساعده هذا التفسير أيضاً في تعديل أوضاعه عند التقدم ثانيةً بطلب آخر للحصول على قرض،

– والبنـــك: بصنع قرار الائتمان استناداً لأحدث طرق التنقيب في البيانات بدلاً من الاعتماد على الأحكام الشخصية والطرق الإحصائية التقليدية.

ويتزامن البحث مع الجهود الحالية للشركة السعودية للمعلومات الائتمانية "سمة" التابعة لمؤسسة النقد العربي السعودي التي تقوم بحساب الجدارة الائتمانية لطالبي القروض من قاعدة بيانات موحدة للقروض الشخصية.

١-٤ أسئلة البحث

١.لماذا تتم الموافقة للبعض على الائتمان في غضون دقائق بعد الانتهاء من تسجيل بياناتهم على الانترنت، بينما تُرفض طلبات البعض الآخر؟ ٢.ولماذا يحصل شخص ما على قرض بمعدل فائدة ١%؟ بينما يُفرض على شخص آخر معدل بأضعاف ذلك؟٣.وما هي أحدث طرق تقييم الائتمان؟وكيف يمكن المقارنة بينها من واقع البيانات التجريبية لتحديد أفضلية إحداها على الطرق الأخرى؟ وكيف تُستخدم لحساب الدرجة الائتمانية لطالب القرض الجديد كمشكلة تصنيف تنبؤية؟٤.وهل تحتوي قاعدة البيانات الموحدة على كافة المتغيرات التي تفيد في تقييم الائتمان وتقدير النموذج؟وما هي أهم المتغيرات الهامة التي يُستند عليها في حساب الدرجة الائتمانية؟

١-٥ أهداف البحث

يهدف البحث إلى:

١.المقارنة بين بعض قواعد بيانات القروض الشخصية الموحدة للتعرف على المتغيرات التي يجب دراستها

٢.التنقيب في بيانات طالبي القروض لنادي الإقراض من خلال طرق الانحدار اللوجستي؛ وشجرة التصنيفات٣. وتقديم نموذج أوتوماتيكي يتم من خلاله التنبؤ بالجدارة الائتمانية لطالب القرض.

٦-١ تنظيم البحث

نُظم البحث في ٧ فصول كالتالي. خُصص الفصل الأول لمقدمة البحث،والثاني لعرض تصميم البحث، والثالث لوصف البيانات والمتغيرات، والرابع لمراجعة الدراسات السابقة،والخامسلعرض المنهجية،والسادسلعرض النتائج التجريبية للبحث، في حين عُرضت الخلاصة والتوصيات في الفصل السابع والأخير.

٢) تصميم البحث

يدرس البحث ويقارن فعالية الانحدار اللوجستي و٣ نماذج لشجرة القرارات (chi squared, entropy reduction, and Gini reduction) للتنبؤ بما إذا كان طالب القرض الجديد سيسدده أم سيتعثر في سداده. وتحتوي مجموعة البيانات على ٥٩٦٠ مقترض، وتهيمن عليها القروض الجيدة في ٤٧٧١ حالة (٨٠%) والقروض السيئة في ١١٨٩ حالة (٢٠%). وقد أُجريت المعاينة الطبقية لتقسيم مجموعة البيانات الكاملة إلى مجموعتي التدريب والاختبار. وقد شملت مجموعة التدريب ٤١٧٠ حالة (70%) منها ٣٣٣٩ قرض جيد و ٨٣١ قرض سيئ. كما شملت مجموعة الاختبار ١٧٩٠ حالة (30%) منها ١٤٣٢ قرض جيد و ٣٥٢ قرض سيئ. وسُجلت النتائج لـ ٣ نقاط فصل احتمالية probability cutoffs: ٠,٣، ٠,٥، ٠,٧. وأظهرت النتائج أن شجرة القرارات قوية وكفؤة في تمييز كل من القروض الجيدة والسيئة في مجموعة الاختبار في ظل اعتبار البنك أن كل القروض الواردة في مجموعة البيانات جيدة وتستدعي تمديد الائتمان. ويقيم البحث ويحلل احتمال التخلف عن السداد للقرض الواحد ولمجموعة من القروض، ويقارن النتائج المتحصل عليها بالنتائج التي نُشرت في الدراسات السابقة.

وتعطي نتائج نموذج الجدارة الائتمانية ما يُعرف باسم بطاقة الدرجة scorecard؛ وهي النتائج من عينة الحاصلين بالفعل على قروض سواء سددوها أم تعثروا في سدادها. ولحساب الدرجة الخاصة بطلب الائتمان الجديد، تُقارن بيانات المتقدم (الجديد) ببطاقة النتائج (للمقترضين القدامى) لتصنيفه في أحد الأنماط السلوكية المشاهدة (سيسدد/سيتعثر) ومن ثم تُحدد درجته التنبؤية. ويخصص نموذج الجدارة الائتمانية درجة لكل خاصية مقاسة لطالب القرض، ثم يتم جمع تلك الدرجات لإنتاج درجة كلية يتقرر بناءً عليها -بالمقارنة مع عتبة threshold معينة يحددها البنك- الموافقة على طلب القرض أو رفضه. وقبل أن يبدأ التحليل، يجب تحويل تكويد المتغيرات متعددة التصنيفات والمتغيرات المستمرة إلى تكويد ثنائي لتسمح بقياس نسب الأرجحية. ثم عمل التحليل الاستكشافي لمتغيرات الدراسة متبوعا بالتحليل متعدد المتغيرات (الانحدار اللوجستي وشجرة التصنيفات). وبعد بناء النموذج (تدريبه) من داخل العينة واختباره وحساب معدل دقة التصنيف الصحيح له، تُختتم العملية باختباره من خارج العينة.

٣) وصف البيانات والمتغيرات

للتنبؤ بصنفي المتغير التابع (تصنيف ما إذا كان طالب الائتمان الجديد: جدير بالائتمان أم غير جدير به)، يجب أولاً تكوين صورة عامة عن المتغيرات المستقلة التي يُحتمل تأثيرها عليه. وقد تم عرض ذلك في القسم الفرعي ١-٢ قبل أن يتمالانتقال لوصف قاعدة بيانات الدراسة في القسم الفرعي ٣-٢ مروراً بفوائد قواعد البيانات الموجودة في القسم الفرعي ٢-٢.

١-٣ التعرف على المتغيرات

تم مراجعة 7 دراسات أُجريت على تقييم ائتمان المستهلك في أوقات مختلفة وأماكن مختلفة (وهو ما يحقق الهدف الأول). ويبين جدول ١ أسماء تلك المتغيرات وفئاتها ومعنوياتها من عدة دراسات سابقة.

جدول ١: المتغيرات التي يُحتمل تأثيرها على الجدارة الائتمانية لقروض المستهلك وفئاتها ومعنويتها

المتغيرات	الفئات	الدراسات			
		A	B	C	D
١- فحص الحساب الجاري	متغير كمي يُحول لمتغير تصنيفي	√*	√*	√*	√*
٢- فترة القرض بالشهور	١٢ أو ٢٤ أو ٣٦ أو ٤٨ أو ٦٠ شهر	√*	√*	√*	√
٣- الالتزام بالسداد (حالة القرض)	متوقف أو متعثر أو متأخر أو جاري أو سُدد بالكامل	√*	√*	√	√
٤- الغرض من القرض	تعزيز المدبونية أو العلاج أو النقل أو التعليم أو ...	√*	√*	√*	√
٥- قيمة الرهن(أو حساب الادخار)	متغير كمي	√	√*	√*	√*
٦- الأصول المتاحة	منزل/ أرض/ سندات أو ودائع مربوطة/ أسهم/ بدون	√*	×	×	×
٧- سنوات الخبرة في الوظيفة الحالية	متغير تصنيفي	√	√*	√	√*
٨- نسبة الخصم من الدخل الخاضع للتصرف	متغير كمي	√	√*	×	×
٩- النوع/الحالة الاجتماعية	متغير تصنيفي	√*	×	√*	×
١٠- مدة الإقامة في المكان الحالى	متغير تصنيفي	√	×	√*	√*
١١- العمر	متغير كمي (تحوله بعض الدراسات إلى تصنيفي)	√*	×	×	×
١٢- وجود ائتمان آخر	متغير تصنيفي	√	√*	×	×
١٣- نوعية السكن	متغير تصنيفي (فيلا-شقة-ايجار)	√	×	√*	√*
١٤- الجنسية	متغير تصنيفي	√*	×	×	×
١٥- عدد الائتمانات السابقة	متغير كمي	√	×	√*	×
١٦- نوعية الوظيفة	متغير تصنيفي	√	×	×	√*
١٧- عدد المعالين	متغير كمي	√	×	√*	√*
١٨- وجود تليفون ثابت	متغير تصنيفي (نعم-لا)	√	×	√*	√*
١٩- معدل الفائدة السنوي	متغير كمي	√*	×	×	×
٢٠- السنوات مع البنك	متغير كمي	×	×	√*	√*
٢١- دخل الزوج الشهري	متغير كمي	×	×	×	√*
٢٢- مبلغ القرض	متغير كمي	√	√	√	√
٢٣- الكفالة	متغير تصنيفي	√	√	√	×
٢٤- دخل طالب القرض الشهري	متغير كمي	×	×	√	√
٢٥- المصاريف الشهرية	متغير كمي	×	×	×	√*
٢٦- رصيد الرهن العقاري غير المسدد	متغير كمي	×	×	×	√
٢٧- يحمل بطاقة ائتمانية	متغير تصنيفي (نعم-لا)	×	×	×	√
٢٨- دخل الأسرة السنوي	متغير كمي	×	×	√	×

- A تشير إلى دراسات: Fahrmeir and Tutz (1994), Giudici (2003), Zurada (2007)
- B تشير إلى دراسة Koh et al. (2006)
- C تشير إلى دراسة Dinh and Kleimeier (2007)
- D تشير إلى دراسة Crook et al. (1992)
- تشير علامة√إلى وجود المتغير في الدراسة، وعلامة × إلى عدم وجوده، وعلامة * إلى معنويته.
- وقد فصل Dinh and Kleimeier (2007) النوع عن الحالة الاجتماعية كل في متغير مستقل، إلا أن دمجهما في متغير تصنيفي واحد (كما هو الحال في معظم الدراسات) هو الأفضل. كما أن مستوى التعليم الذي ظهر في نفس الدراسة (كمؤشر للدخل) يُعد مرادفاً لنوعية الوظيفه في الدراسات الأخرى. ويُفضل أن يعتمد البنك على متغير نوعية الوظيفة (كمؤشر أفضل للدخل) بدلاً من اعتماده على متغير مستوى التعليم، لأنه من الممكن أن يحمل شخص درجة الدكتوراه وهو عاطل عن العمل. وأهمل البحث متغير حساب الادخار (الودائع المربوطة) لأن ذلك أمر محدود للغاية في المجتمع السعودي.

ويتضح من العرض السابق أن دراسات تقييم الائتمان قد اتفقت على معنوية تأثير الخمس متغيرات الأولى فقط، لكنها لم تتفق بشأن أسماء وعدد المتغيرات المستقلة التي يُحتمل تأثيرها على متغير حالة القرض. وقد كان ذلك طبيعياً لأن تلك المتغيرات تتوقف بدورها على الجوانب السلوكية والظروف الاقتصادية للبلاد التي أُجريت فيها تلك الدراسات والقيود الائتمانية المفروضة. ويفيد جدول ١ في التعرف على المتغيرات الأكثر استخداماً في البنوك التي تطبق نماذج تقييم الائتمان وهو ما يخدم إنشاء قاعدة بيانات موحدة (أو تعديل قاعدة البيانات الحالية).

٢-٣ منافع قواعد البيانات الموحدة

يحقق وجود قاعدة بيانات موحدة المنافع التالية:

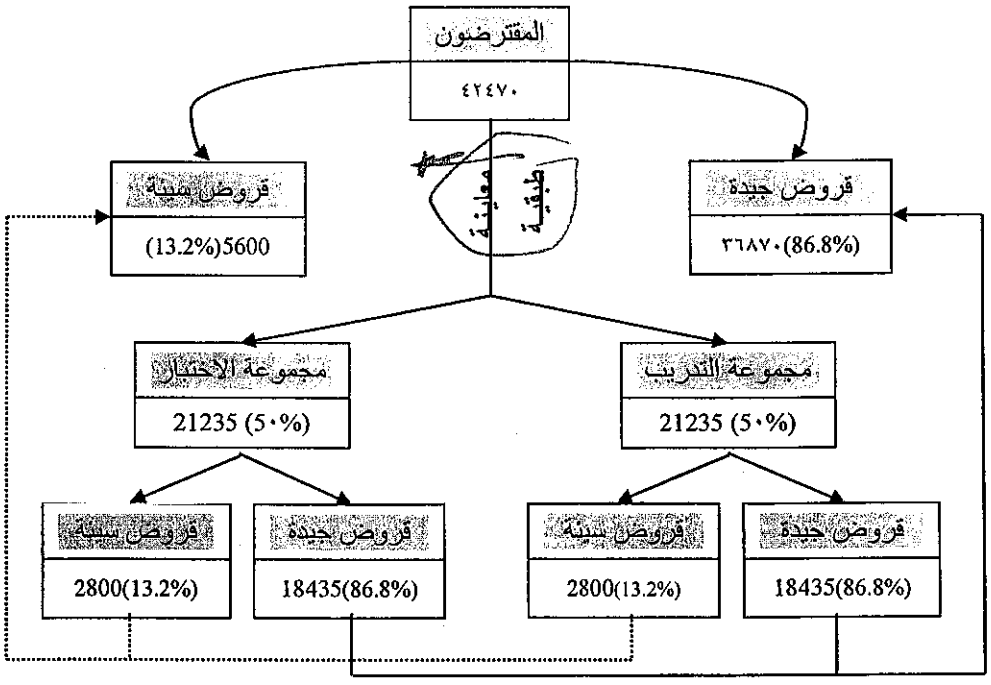
الحدمن مشكلة القيم المفقودة: يرجع وجود القيم المفقودة إلى عدم كفاءة نظام المعلومات ببعض الفروع، و/أو إهمال بعض الموظفين في إدخال تلك القيم للنظام، و/أو عدم رغبة بعض المقترضين في الإجابة عن الأسئلة المتعلقة بتلك القيم. ويقوم بعض الباحثين بحل تلك المشكلة بحذف السجلات التي تحتوي على تلك القيم وهو ما يؤدي لاختزال حجم العينة والتوصل لنتائج غير موثوق فيها، أو باستبعاد كافة الحقول (المتغيرات) التي تحتوي على القيم المفقودة وهو ما يؤدي لنقص دقة توصيف النموذج. إلا أن وجود قاعدة بيانات موحدة حديثة -يتم التسجيل فيها عبر الانترنت وتشترط ملء جميع الحقول لإكمال التسجيل- يقضي على هذه المشكلة تماماً.

* *القضاء على مشكلة التحيز*: تنتج مشكلة التحيز في قواعد البيانات القديمة (السجلات الخاصة بالبنوك) بسبب عدم تمثيل العينة للمجتمع لاحتواءها على بيانات القروض المقبولة فقط دون المرفوضة. وفي حالة وجود قاعدة بيانات موحدة، فإن بيانات طالبي القروض الذين رُفضت طلباتهم سوف تكون متاحة للدراسات القادمة إلى جانب بيانات القروض المقبولة وهو ما يقضي على مشكلة التحيز.

* *وضع درجات مدروسة لسمات طالبي القروض*: بدلاً من تصنيف كل بنك لتلك السمات وإعطائها درجات بالطريقة التي يراها مناسبة.

٣-٣ قاعدة بيانات نادي الإقراض

تحتوي قاعدة البيانات الخام لنادي الإقراض (خلال فترة الدراسة) على ٤٢٥٣٩ حالة و ١٠١ متغير. وبعد استبعاد السجلات والحقول ذات القيم المفقودة والمتغيرات التعريفية (كالعنوان ورقم تعريف العميل)، فقد قُلص حجم مجموعة البيانات إلى ٤٢٤٧٠ حالة مسجلة على 22 متغير. وقد استُخدم متغير حالة القرض الذي يحتوي على ٦ تصنيفات -قروض مشطوبة charged off (11.1%)، ومتعثرة (1%.)default، وفي فترة السماح (4%.)، ومتأخرة أقل من شهر (3%.)، ومتأخرة أكثر من شهر (1.3%)، وجارية(38.5%)، ومدفوعة بالكامل(48.3%)- لتصنيف القروض إلى جيدة وسيئة. وكما هو مبين في شكل ١ وجدولي ٢ و ٣، فقد صُنف القرض على أنه سيء إذا وقع في أحد التصنيفات الأربعة الأولى (5600 حالة بنسبة ١٣,٢%)، كما صُنف القرض على أنه جيد إذا وقع في أحد التصنيفين الأخيرين (٣٦٨٧٠ حالة بنسبة ٨٦,٨%). وبمعنى آخر، فقد حُول متغير حالة القرض في جدول ٢ إلى متغير ثنائي في جدول ٣؛ ليصبح الأخير هو المتغير التابع للدراسة. وقد طُبقت المعاينة الطبقية وخُصص 50% من العينة الكلية لمجموعة التدريب والـ 50% الباقية لمجموعة الاختبار. وقُسّمت مجموعتي التدريب والاختبار أيضاً إلىقروض جيدة وقروض سيئة.وقد بُنيت النماذج من مجموعة التدريب، واختُبر أدائها من مجموعة الاختبار.



شكل (١): وصف خطة الدراسة

وباستخدام مجموعة البيانات هذه، فقد تم بناء3نماذج للانحدار اللوجستي و٣ نماذج لشجرة القرارات؛ للتنبؤ بما إذا كان مقدم طلب الائتمان الجديد سيتعثر في سداد قرضه أم سيسدده. وقد شملت الدراسةمتغير تابع واحد و٢٢ متغير مستقل، وقد استُخدم متغير (حالة القرض) في اشتقاق صنفي المتغير التابع ليصبح عدد المتغيرات المستقلة ٢١ مقابل متغير تابع واحد تصنيفي ثنائي. ولم يمكن دراسة أثر بعض الخصائص الشصية مثل: النوع والحالة الاجتماعية ومستوى التعليم والعمر والجنسية -على الرغم من أهميتها المحتملة في المنطقة العربية- على الجدارة الائتمانية لقروض المستهلكين لعدم وجودها في قاعدة البيانات.

المتغيرات المستقلة ٢١ مقابل متغير تابع واحد تصنيفي ثنائي. ولم يمكن دراسة أثر بعض الخصائص الشصية مثل: النوع والحالة الاجتماعية ومستوى التعليم والعمر والجنسية -على الرغم من أهميتها المحتملة في المنطقة العربية- على الجدارة الائتمانية لقروض المستهلكين لعدم وجودها في قاعدة البيانات.

٤) الدراسات السابقة

كان طبيعياً أن تركز الدراسات الأولى لتقييم الائتمان (مع بداية التسعينات) على تأسيس قواعد البيانات وتحديد المتغيرات المؤثرة على الجدارة الائتمانية باستخدام الطرق الإحصائية التقليدية للتحليل متعدد المتغيرات.فقد اعتمدت بعض هذه الدراسات في تقييمها للائتمان على تحليل التمايز الذي يتطلب فروضاً إحصائية يصعب قبولها في الواقع العملي. لذلك فقد استبدلت الدراسات التالية تحليل التمايز بالانحدار اللوجستي لأنه أقل تقييداً. كمااعتمدت هذه الدراسات أيضاً -ولكن بدرجة أقل- على الخوارزمية الجينية وطريقة أقرب الجيران والبرمجة الخطية والأنظمة الخبيرة Thomas (2000).ومن أقدم وأشهر الدراساتالتي تعاملت مع موضوع الجدارة الائتمانية دراسة Crook et al. (1992)،ودراسة Fahrmeir and Tutz (1994). ولكن الدراسات الأحدث طبقت الطرق الحديثة للتنقيب في البيانات (وبالتحديد أساليب تعليم الآلة machine learning) على مشكلتي تقييم الائتمان وتقييم المخاطر الائتمانية. وكان مدخل شجرة القرارات هو الأسلوب الأكثر استخداماً لبناء نماذج تقييم ائتمان المستهلك.

وقد درس Crook et al. (1992) تأثير ١٥ متغير على الجدارة الائتمانية لطالبي القروض الشخصية في المملكة المتحدة باستخدام تحليل التمايز، وتبين أن ١٢ منها تؤثر معنوياً على متغير حالة القرض (جدول ١). كما قيم Dinh and Kleimeier (2007) الائتمان في فيتنام بإدخال ٢٢ متغير مبدئي إلى نموذج الانحدار اللوجستي، ليظهر أن ١٦ منها تؤثر معنوياً على متغير حالة القرض (جدول ١ تحت العمود C)، منها ٧ تتفق -وإن كانت بأهمية ترتيب مختلفة- مع دراسة Crook et al. (1992).

كماأسست دراسة Fahrmeir and Tutz (1994) قاعدة بيانات لأحد البنوك الألمانية الكبرى تحتوي على متغير تابع، ٢٢ متغير مستقل. واعتمدت عليها بحوث تطبيقية كثيرة (منها Giudici, 2003; Koh et al., 2006; Zurada, 2007) في تقديمها لطرق تقييم الائتمان. إذ قام Giudici (2003) بعمل التحليل الاستكشافي وبناء نماذج الانحدار اللوجستي وشجرة التصنيفات. وقام Koh et al. (2006)بتقديم طريقة من خطوتين لبناء نموذج مهجن يجمع بين نماذج الانحدار اللوجستي وشجرة التصنيفات والشبكات العصبية باستخدام Clementine (أحد برامج التنقيب في البيانات). واستخدم Zurada (2007) مدخل شجرة القرارت لتقدير ٣ نماذج مختلفة لها، وبين أن أفضل نموذج هو الذي يعتمد على طريقة chisquared.

وطبقت دراسة Feldman and Gross (2005) أساليب شجرة القرارات للكشف عن معدلات التخلف عن سداد الرهن العقاري. وقامت دراسة Kamleitner and Kirchler (2007) بمراجعة الأبحاث السابقة المتعلقة بعملية الجدارة الائتمانية للمستهلك من ٣ نواحي: قبل منح الائتمان، وعند منح الائتمان، وبعد منح الائتمان. كما قارنت دراسة Sabzevari, Soleymani and Noorbakhsh بين الطرق الإحصائية التقليدية (probit and logistic regression) من ناحية وطرق التنقيب في البيانات (CART and MARS) من ناحية أخرى بهدف فحص أداء النماذج المختلفة للجدارة الائتمانية باستخدام بيانات ٣٧ متغير.

ولم يعثر الباحث على أي بحوث إحصائية تخص مشكلة تقييم الائتمان في المنطقة العربية في أي من قواعدالبحوث على الانترنت، وهو ما يضفي على هذا البحث أهمية خاصة. كما يتميز البحث بتجنبه لمشكلة التحيز الذي وقعت فيه معظم دراسات تقييم الائتمان –بسبب اعتمادها على بيانات القروض المقبولة فقط دون المرفوضة،حيثتتيح قاعدة البيانات منح الائتمان أيضاً لمن لم تتمشى خصائصهم مع السياسة الائتمانية للبنك.

٥) منهجية البحث

يُعد الانحدار اللوجستي وشجرة التصنيفات (استقراء القاعدة) من أشهر وأحدث طرق التنقيب في البيانات عند دراسة مشكلة تقييم الائتمان بغرض تصنيف طالب القرض الجديد في أحد فئتين: جدير بالائتمان، أوغير جدير. ويجيب هذا القسم على سؤال البحث الثالث.

١-٥الانحدار اللوجستي

إستناداً إلى العديد من البحوث التي راجعها Thomas (2000)، فقد كان الانحدار اللوجستي أكثر الطرق استخداماً لتقييم الجدارة الائتمانية لقروض المستهلك من بين طرق الانحدار الخطي المتعدد وتحليل التمايز والخوارزمية الجينية وطريقة أقرب الجيران والبرمجة الخطية والأنظمة الخبيرة. وقد تم التعامل مع الانحدار اللوجستي في هذا البحث وفقاً للمنهجية التالية:

١- **تحديد المتغيرات وتكويدها:** يبدأ تحليل الانحدار اللوجستي بتحديد المتغير التابع والمتغيرات المستقلة. ولا يوجد خلاف في دراسات تقييم القروض الشخصية على أن المتغير التابع هو "حالة القرض" الذي يأخذ شكلاً ثنائياً (يسدد=١/متعثر=٠)، حيث تُحسب حالات التعثر بأثر رجعي *ex-post* من خلال عينة المقترضين الفعليين بمراجعة التزامهم بالسداد. وكما هو مبين بجدول ٤، فقد تم تكويد المتغيرات التصنيفية بإعطاء فئاتها درجات من الأعلى(حسب عدد فئات المتغير) إلى الأدنى (الدرجة ٠)، مع تخصيص الدرجة الأعلى للفئة ذات الجدارة الأعلى (التي تحتوي على أقل عدد من المتعثرين). ويتم التوصل في هذه المرحلة المبكرة لقائمة متغيرات مبدئية؛ حتى مع البنوك التي لا تطبق نماذج تقييم الائتمان، لأن تلك البنوك لن تتمكن من صنع قرار الائتمان بدون جمع معلومات عن طالبي القروض. غير أن المشكلتين الحقيقيتين اللتان قد تعيقا تطبيق الانحدار اللوجستي – في ظل عدم وجود قاعدة بيانات موحدة– هما وجود كم كبير من القيم المفقودة في معظم المتغيرات و/أو وجود التحيز.

٢- ***النموذج:*** ولكن احتمال التعثر في السداد لطالب القرض الجديد يكون غير قابل للمشاهدة قبل حدوثه *ex-ante*، ويستطيع أسلوب الانحدار اللوجستي أن يتغلب على ذلك بتقدير هذا الاحتمال π_j لكل طالب قرض جديد باستخدام طريقة الامكان الأكبر التكرارية iterative maximum likelihood كالتالي:

$$\pi_j = \frac{1}{1+e^{-Z_j}} \qquad (1)$$

حيث تشير Z_j إلى الدرجات (توليفة خطية للمتغيرات المستقلة المُعرّفة لخصائص طالب القرض مرجحة بالمعاملات β_k) حسب المعادلة التالية:

$$\hat{z_j} = \beta' x = \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} \qquad (2)$$

ووفقاً للمعادلتين السابقتين، فإن الدرجات الكبيرة تعطي احتمالات تعثر ضئيلة.

٣- ***اختيار المتغيرات وتقدير النموذج:*** يتم إدخال جميع المتغيرات المبدئية للمعادلة ١ وتصفيتها للحصول على أقصى دقة تنبؤية للنموذج بتطبيق الانحدار المتدرج للأمام forward stepwise (بإضافة المتغيرات للنموذج واحداً تلو الآخر) والتأكد من هذا الاختيار بتطبيق الانحدار المتدرج للوراء backward stepwise (بحذف المتغيرات من النموذج واحداً تلو الآخر). وتنتهي هذه المرحلة بتحديد المتغيرات الهامة والتوصل لتقديرات لمعاملاتهاوحساب الدرجة التي حصل عليها طالب القرض الجديد وفقاً للمعادلة ٢.

٤- ***اختبار الدقة التنبؤية:*** وتُختتم العملية باختبار الدقة التنبؤية للنظام، ويُفضل أن يكون ذلك من خارج العينة out-of-sample لأن التنبؤات من داخل العينة in-sample تعطي مستويات دقة مبالغ فيها. ويتم ذلك بمقارنة الوضع المشاهد بالمتوقع باستخدام مصفوفة الالتباس confusion matrix المبينة بجدول ٢.

جدول ٢: مصفوفة التباس نظرية

المتوقع / المشاهد	غير متعثر (جيد)	متعثر (سيئ)	نسبة التصنيف
غير متعثر (جيد)	G_g	G_b (خطأ النوع الأول)	للقروض الجيدة $G_g/(G_g+G_b)$
متعثر (سيئ)	B_g (خطأ النوع الثاني)	B_b	للقروض السيئة $B_b/(B_b+B_g)$
الحساسية sensitivity والتحديد specivity	$G_g/(G_g+B_g)$	$B_b/(B_b+G_b)$	الكلي $(G_g+B_b)/(G_g+G_b+B_g+B_b)$

وتقارن المصفوفة التصنيفين المشاهد والمتوقع من خلال ٤ تصنيفات:

- إثنان صحيحان على القطر الرئيسي: G_g (تصنيف النموذج للقرض على أنه جيد وهو جيد بالفعل)، B_b (تصنيف النموذج للقرض على أنه سيء وهو سيء بالفعل)،
- واثنان خاطئان على القطر الثانوي: إذ يشير G_b إلى عدد القروض الجيدة التي صنفها النموذج (بالخطأ) على أنها سيئة أو الخطأ من النوع الأول، كما يشير B_g إلى عدد القروض السيئة التي صنفها النموذج (بالخطأ) على أنها جيدة أو خطأ النوع الثاني،
- ويعطي الجدول ٢ نسب تُستخدم لمقارنة آداء النماذج، هي: نسبة التصنيف الصحيح للقروض الجيدة (عدد القروض الجيدة التي صنفها النموذج على أنها جيدة/عدد القروض الجيدة)، ونسبة التصنيف الصحيح للقروض السيئة (عدد القروض السيئة التي صنفها النموذج على أنها سيئة/عدد القروض السيئة)، ونسبة التصنيف الصحيح الكلي (وهي مجموع التصنيفين الصحيحين إلى إجمالي عدد القروض) ويُعرف مكمل هذه النسبة بمعدل سؤ التصنيف، والحساسية (نسبة القروض التي صنفها النموذج على أنها جيدة وهي جيدة إلى إجمالي عدد التصنيفات الجيدة للنموذج)، والتحديد (نسبة القروض التي صنفها النموذج على أنها سيئة وهي سيئة إلى إجمالي عدد التصنيفات السيئة للنموذج).

٤-٥ شجرة القرارات

بينما ينتج الانحدار اللوجستي الدرجات أولاً ثم يصنفها استناداً لقاعدة تمايز ما، فإن نماذج شجرة القرارات تبدأ بتصنيف المشاهدات في مجموعات ثم انتاج درجة كل مجموعة (Giudici, 2003, p. 100). وتنقسم نماذج الشجرة إلى: *أشجار الانحدار* (ويكون فيها المتغير التابع كمي)و*أشجار التصنيف* (ويكون فيها المتغير التابع تصنيفي) وهي حالة هذا البحث.

ويُعد مُصنِف شجرة القرارت من أبسط أساليب التصنيف، كما أنه من أكثرها قبولاً وأوسعها (وأحدثها) استخداماً بالنسبة لمشاكل تقييم الائتمان. إذ أن تحديد شجرة القرارات للمتغيرات الأكثر أهمية في التأثير على المتغير التابع (تقييم الائتمان)، يمكن الباحثين من استخدامها كطريقة لاختيار المتغيرات إلى جانب استخدامها كطريقة لبناء النماذج (Linoff and Berry, 2011, p. 199). ولشجرة القرارات ٣ أنواع من العُقَد nodes: عقدة الجذر، والعقدالداخلية، والعقد الخارجية (الأطراف أو الأوراق). وتكون عقدة الجذر في الأعلى بلا حواف داخلة، والأوراق بالأسفل بلا حواف خارجة. وفي الشجرة الثنائية، يخرج من عقدة الجذر فرعانيوصلان إلى عقدتين داخليتين، ويتفرع من كل عقدة فرعين إلى أن نصل للعقد الطرفية. فكل عقدة داخلية يدخل عليها فرع واحد ويخرج منها فرعين، وكل عقدة خارجية يدخل عليها أيضاً فرع واحد ولكن لا يخرج منها أية فروع. وتمثل كل عقدة طرفية عنوان فئة ما، أما الفروع فتحتوي على حالات اختبار السمة التي تنقسم لحالات منفصلة بخصائص مختلفة. وتوظف الخوارزميات الكفؤة استراتيجيات البحث للحصول على دقة معقولة، بحيث يتم التوصل للحجم الأمثل للشجرة بلا نمو زائد أو ناقص بأسرع ما يمكن.

وتركز خوارزميات بناء شجرة القرارت على معالجة مسألتين: إنقسام أنماط التدريب، ومتى ينبغي وقف هذا التقسيم. ويُبنى أفضل انقسام على درجة الاضطراب أو عدم النقاوة disorder/impurity المصاحبة للعقد الداخلية. ومن أمثلة مقاييس الاضطراب (عدم الانسجام):

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i \setminus t) \log_2 p(i \setminus t) \qquad (3)$$

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i \setminus t)]^2 \qquad (4)$$

$$Classification\ error(t) = 1 - \max_i [p(i \setminus t)] \qquad (5)$$

حيث تشير c إلى عدد الفئات، $0 \log_2 0 = 0$في حسابات الإنتروبي، $p(i \setminus t)$ إلى كسر الحالات التي تنتمي للفئة iعند العقدة t. وإذا حُذف الرمز t، فإن الكسر يُبسط إلى p_i. ويُعَبر عن توزيع الفئات في الشجرة الثنائية بالكسرين (p_0, p_1) حيث $p_1 = 1 - p_0$.

وإذا تم التوصل للتقسيم الأمثل النهائي، فإن شجرة الانحدار تُنتج قيمة مقدرة $\hat{y}_i$ (لكل مشاهدة في المتغير التابع y_i تساوي متوسط قيم المتغير التابع في المجموعة التي تنتمي إليها المشاهدة رقم i:

$$\hat{y}_i = \frac{1}{n_m} \sum_{l=1}^{n_m} y_{lm} \qquad (6)$$

حيث تشير m إلى رقم المجموعة التي يُحسب لها القيمة الموفقة، وتشير n_m إلى حجمها. أما في حالة شجرة التصنيف، فتُحسب تلك القيم بمعلومية الاحتمالات المقدرة لانتساب المشاهدة لمجموعة معينة. ويكون احتمال النجاح في حالة التصنيف الثنائي:

$$\pi_i = \frac{1}{n_m}\sum_{l=1}^{n_m} y_{lm} \qquad (7)$$

وتأخذ المشاهدة y_{lm} القيمة 0 أو 1، لذلك فإن الاحتمال المقدر يناظر نسبة النجاح في المجموعة m. ومن الجدير بالذكر أن كلا من $\hat{y}_i$ ، π_i يُعد ثابتاً لكل المشاهدات.

ولمعرفة مدى جودة شرط الاختبار، تُقارن درجة الاضطراب في العقدة الأصل parent node بدرجة الاضطراب في أبنائها child nodes (العقد المتفرعة منها)، وكلما ارتفعت قيمة المكسب Δ ، كلما كان الانقسام أفضل.

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j) \qquad (8)$$

حيث تشير $I(.)$ إلى مقياس الاضطراب لعقدة ما، N إلى عدد الحالات الكلية في العقدة الأصل، k إلى عدد قيم الصفة، $N(v_j)$ إلى عدد الحالات المقترنة بالعقدة الطفل v_j.

وقد تم بناء شجرة القرارات في هذا البحث (أو السماح لها بالنمو) بثلاث طرق:

١. خوارزمية CHAID: أي اكتشاف الفاعل التلقائي باستخدام مربع كاي Chi-squared Automatic Interaction Detection. وتختار هذه الخوارزمية المتغيرات المستقلة صاحبة أقوى تفاعل مع المتغير التابع، كما تدمج الفئات غير المعنوية في كل متغير مستقل.

٢. خوارزمية CRT: أي أشجار التصنيف والانحدار Classification and Regression Trees. ويتم فيها تقسيم البيانات إلى شرائح متجانسة بقدر الإمكان بالنسبة للمتغير التابع، بحيث تحتوي العقدة الطرفية على نفس القيمة للمتغير التابع.

٣. خوارزمية QUEST: أي الشجرة الإحصائية السريعة غير المتحيزة الكفؤة Quick, Unbiased, Efficient Statistical Tree. وهي تتجنب تحيز الطرق الأخرى للمتغيرات المستقلة صاحبة أكبر عدد من الفئات.

وتركز معايير التقسيم المبنية على Chi-squared على تقليل التغير في التوزيع المستهدف للعقد الفروع (الأبناء). إذ تُعد إحصاءة الاختبار (likelihood ratio Pearson chi-squared test statistic) مقياساً للاقتران بين فئات المتغير التابع والعقد الفروع. ويمكن استخدام هذه الإحصاءة في الحكم على جدارة التقسيم بقياس الفرق بين تكرارات الخلايا المشاهدة والمتوقعة بفرض استقلال عقد الفروع عن العقد المستهدفة. وقد افتُرض في هذا البحث أن مستوى المعنوية يساوي ٠,٠٥. ولمزيد من التفاصيل حول شجرة القرارات، يمكن الرجوع إلى واحد أو أكثرمن المراجع التالية: Linoff and Berry, 2011; Giudici, 2003; dtreg Sofware: http://www.dtreg.com; andSAS Enterprise Miner: http://www.sas.com

٦) النتائج التجريبية

يدرس البحث أثر ٢١ متغير مستقل (١٦ كمي، ٥ تصنيفي) على متغير تابع واحد تصنيفيy هو حالة القرض. ووفقاً لمنهجية البحث، فقد بدأ التحليل بعمل التحليل الاستكشافي بغرضي دراسة أثر كل متغير مستقل على المتغير التابع من خلال التحليل ثنائي المتغيرات وتصفية مجموعة البيانات باستبعاد السجلات ذات القيم المتطرفة والمفقودة. ثم تماختيار المتغيرات المعنوية وادخالها لنموذج التحليل متعدد المتغيرات دفعة واحدة.

٦-١ التحليل الاستكشافي

أُجري التحليل الاستكشافي باستخدام التحليل أحادي المتغيرات لتنقيح البيانات ووصف أنماط الدراسة. كما أُجري التحليل ثنائي المتغيراتلبحث شدة الروابط الممكنة بين كل متغير مفسر والمتغير التابع وبهدف اختيار المتغيرات.

٦-١-١ التحليل أحادي المتغيرات

- **مبلغ القرض، واجمالي الدفعات، واجمالي المسدد من أصل القرض**

تتراوح مبالغ القروض الشخصية في مجموعة بيانات الدراسة بين ٥٠٠$و٣٥٠٠٠$بمنوال ١٠٠٠٠$بإجمالي٤٧١مليوندولار. وقد حُذفت سجلات القروض الكبيرة (التي تزيد عن٣٥٠٠٠$) من التحليل لكونها تخرج عن دائرة القروض الشخصية ولأن عددها محدود ولأن مبالغها تمثل قيم شاذة.ويصعب الحكم على مجمل نشاط البنك في فترة الدراسة؛ حيث أن المبالغ المسددة بالكامل لم تتجاوز ٤٨,٣%، في حين أن٣٨,٥% من القروض لا زال جارياً، ١٣,٢% من القروض التي استُحقت آجالها صُنفت على أنها متعثرة.وبدراسة تلك المبالغ للقروض المتعثرة ومقارنتها مع نظائرها في العينة الكلية، اتضح أن إجمالي القروضالخاصةبالمتعثرين بلغ64,369,450$ بنسبة ١٣,٧% من إجمالى مبالغ االقروض، سُدد منها15,389,256$؛ لتكون نسبة الخسائر ٧٦% (من إجمالي أصول القروض المتعثرة) و 40% (من إجمالي أصول القروض ككل).ولم يكن ذلك بسبب حجم البيانات المفقودة الذي ظهر أمام متغير المسدد من أصل القرض في الجدولين الأولين بالملحق، لأن النسبتين السابقتين لم تتغيرا بشكل جوهري بعد حذف السجلات الخاصة بتلك البيانات المفقودة (أصبحتا ٧٢,٤%، ٣١% على الترتيب). وباستكشاف متغير المبلغ المسدد من الأصل –بعد تحويله إلى الشكل التصنيفي على النحو المبين بجدول ٤– بحسب حالة السداد لمجموعة البيانات ككل، اتضح أن التعثر يتناسب عكسياً مع مبلغ القرض؛ فهو أكبر للقروض الصغيرة عنه في القروض المتوسطة ثم الكبيرة.

- **الدخل السنوي، ونسبة الدين إلى الدخل، والتحقق من الدخل**

يحتل الدخل السنوي للمقترض/طالب القرض أهمية كبيرة في دراسات تقييم الائتمان. إذ يُعد ذلك الدخل إجازة لمنح القرض لطالبه، ومؤشراً لاحتمال السداد المتوقع للمقترض. ويتراوح الدخل السنوي لأفراد عينة الدراسة بين1,896$ و 750,000$ بمتوسط 69,594$. وقد تم حذف بيانات ١٦ حالة عند هذه المرحلة بسبب دخولهم المتطرفة التي تجاوز المليارات –لأنه حتى إذا كانت تلك الدخول صحيحة، فإن قيمها ستؤثر على دقة التحليل. ويُعد أيضاً نسبة الدين إلى الدخل من المؤشرات الهامة لمنح الائتمان؛ أي نسبة الدفعة الشهرية إلى الدخل الشهري الخاضع للتصرف. إذ كلما انخفضت تلك النسبة، كلما شكّل ذلك دافعاً جيداً لمنح الائتمان. وقد تراوحت نسبة الدين إلى الدخل بين ٠% إلى ٣٠% لكل من العينة الكلية وعينة المتعثرين على حد سواء.

وبلغت نسبة الذين لم يتحقق البنك من دخولهم السنوية ٦٨,٣%، ثلثاهما من القروض السيئة؛ وهو ما يفسر سبب الخسائر الكبيرة التي تعرض لها البنك.

- **معدل الفائدة، والقسط الشهري، وأجل القرض، والتصنيف الائتماني، والسياسة الائتمانية**

يُفرض معدل الفائدة على القرض حسب تصنيفه الائتماني (الدرجة الائتمانية) وأجله (مدته) ومبلغ القسط ومدى التوافق بين خصائص العميل والسياسة الائتمانية التي يفرضها البنك. وعند استكشاف بيانات متغير القسط الشهري، تبين أن هناك ٧ قيم متطرفة (تتجاوز مبلغ القرض). وقد تم استبعاد السجلات الخاصة بتلك الحالات حتى لا تؤثر على دقة التحليل.

وقد تراوحالقسط الشهري الذي يدفعه المقترض بين 15.7$و1305.2$ بمتوسط 322.5$ (بعد أن كان 3008$ قبل الحذف)، كما تراوح معدل الفائدة بين ٥,٤% و٢٤,٦%. وبلغت القروض متوسطة الأجل لمدة ٣ سنوات ٧٥% من حجم العينة، والباقي كان لمدة ٥ سنوات. وكانمتوسط معدل الفائدةلمدة ٣ سنوات ١١,٢%، وارتفع متوسطه بزيادة أجله لـ٥ سنوات إلى١٤,٩%.

وتوافق معظم البنوك –إجابةً على سؤال البحث الأول– على تلبية القروض للعملاء ذوي الدرجات الائتمانية العليا. وبالتالي فإن بياناتها تقتصر على خصائص هذه الفئة فقط، وهو ما يعرض نتائج نماذجها لمشكلة التحيز الناجمة عن عدم شمول خصائص العملاء ذوي القروض المرفوضة في البيانات. وتخالف مجموعة بيانات الدراسة هذا الشرط وتلبي كافة الطلبات، وهو ما يغلبها على مشكلة التحيز المذكورة. ويتناسب معدل الفائدة عكسياً مع التصنيف الائتماني للعميل؛ إذ بلغ متوسطه للتصنيف Aحوالي ٧,٣% وارتفع تدريجياً بانخفاض التصنيف إلى أن وصل إلى ٢٠,٤% للتصنيف G. كما يتوقف معدل الفائدة أيضاً على مدى مطابقة خصائص العملاء للسياسة الائتمانية للبنك؛ حيث بلغ ١٤,٢% في المتوسط لمن لم تنطبق عليهم الشروط مقابل ١٢% لمن انطبقت عليهم الشروط.وكنتيجة لتفاعل العوامل السابقة، كان من الطبيعي أن يكون معدل الفائدة للقروض السيئة أعلى منه للقروض الجيدة؛ حيث بلغ متوسطاهما ١٣,٨% مقابل ١١,٩%على الترتيب.

والخلاصة، أن معدل الفائدة يرتفع بانخفاض التصنيف الائتماني، وعدم مطابقة خصائص العملاء مع السياسة الائتمانية للبنك،وطول أجل القرض، وكبر مبلغ القرض وبالتالي مبلغ القسط؛ وهو ما يجيب على سؤال البحث الثاني.

- **الخبرة الوظيفية، وامتلاك منزل، وغرض القرض**

شكّل طرفا الخبرة الوظيفية (الذين زادت خبرتهم عن ١٠ سنوات، والذين قلت خبرتهم عن سنة) أكثر الفئات التي استفادت من القروض (٤٢,٣%)، وبلغت نسبتهما من القروض السيئة ٤٣,٧%. كما مثلتا فئتا المستأجرون والرهن العقاري ٩٢% من الحاصلين على القروض في متغير امتلاك المنزل بنسبة ٩٢% أيضاًمن القروض السيئة. أما بالنسبة لمتغير غرض القرض، فقد كان تعزيز المديونية أكثر الفئات خطراً على الجدارة الائتمانية؛ حيث مثلت هذه الفئة بمفردها ٤٣,٨%.

- **عدد الاستعلامات، وعدد السجلات العامة المهينة، وعدد خطوط الائتمان الجارية والكلية، والرصيد المتجدد، ومعدل الاستفادة من الخط المتجدد**

على الرغم من تراوح عدد الاستعلامات عن المقترضين بين ٠ إلى ٣٣، إلا أنه لم يتجاوز ٣ في ٩٥% من المقترضين. كما لم يتجاوز عدد السجلات العامة المهينة number of derogatory public records العدد ثلاثة، مع ملاحظة

عدم وجود أي سجلات مهينة في ٩٤,٤% من العينة ككل. وتراوح عدد خطوط الائتمان الجارية بين ١ و ٤٧ خط بمتوسط ٩ خطوط،والكلية بين ١ و ٦٦ للقروض السيئة وبين ١ و ٩٠ للقروض الجيدة بمتوسط ٢٢ للحالتين. كما تراوح رصيد خط الائتمان المتجدد total credit revolving balanceبين ٠ و 1207359$للقروض السيئة بمتوسط 15261$، وكان العبء أقل بالنسبة للقروض الجيدة؛ حيث تراوح الرصيد بين ٠ و 602519$بمتوسط14035$. وأخيراً، يقيس معدل الاستفادة من الخط المتجدد نسبة مبلغ القرض الحالي إلى إجمالي المبالغ المقترضة من كافة القروض، وقد بلغ متوسط هذا المعدل ٥٦% في القروض السيئة و ٤٨% للقروض الجيدة.

٦-١-٢ التحليل ثنائي المتغيرات

يتيح التحليل ثنائي المتغيرات معرفة كفاءة كل متغير مفسر في تحديد العملاء غير الموثوق فيهم (Y=0)، وهو ما يفيد في اختيار المتغيرات فيما بعد في مرحلة التحليل متعدد المتغيرات. وقد تم عمل ذلك باستخدام نسب الأرجحية odds ratios المبينة نتائجها في جدول ٣ التالي:

جدول ٣: نسب الأرجحية للاقتران بين كل متغير مفسر وحالة القرض

No.	Variable	type	Wald	df	Sig.	Odds ratios	95% C.I.for EXP(B)	
							Lower	Upper
1	installment	continuous	1563.926	1	.000	0.975	.973	.976
2	total_pymnt	continuous	638.217	1	.000	1.001	1.001	1.001
3	grade	categorical	580.164	6	.000			
	grade(1)		379.473	1	.000	0.001	.001	.003
	grade(2)		420.231	1	.000	0.003	.002	.005
	grade(3)		507.085	1	.000	0.006	.004	.009
	grade(4)		567.103	1	.000	0.011	.007	.016
	grade(5)		531.629	1	.000	0.033	.025	.044
	grade(6)		415.111	1	.000	0.122	.100	.149
4	total_rec_prncp	continuous	529.789	1	.000	1.001	1.001	1.001
5	int_rate	continuous	479.268	1	.000	1.563	1.502	1.627
6	purpose	categorical	108.650	13	.000			
7	loan_amnt	continuous	51.662	1	.000	1.000	1.000	1.000
8	amnt	categorical	67.031	2	.000			
	amnt(1)		56.113	1	.000	0.236	0.162	0.344
	amnt(2)		66.931	1	.000	0.322	0.246	0.423
9	inq_last_6mths	continuous	51.099	1	.000	.889	.861	.918
10	pub_rec	continuous	28.331	1	.000	.662	.569	.771
11	revol_bal	continuous	28.001	1	.000	1.000	1.000	1.000
12	annual_inc	continuous	15.933	1	.000	1.000	1.000	1.000
13	credit_policy(1)	continuous	8.584	1	.003	.732	.594	.902
14	total_acc	continuous	5.703	1	.017	.990	.982	.998
15	open_acc	continuous	4.219	1	.040	1.019	1.001	1.037
16	is_inc_v(1)	continuous	4.617	1	.032	.891	.801	.990
17	term(1)	continuous	9.909	1	.002	.774	.660	.908
18	emp_length	categorical	17.768	10	.059			
19	dti	continuous	3.369	1	.066	1.007	1.000	1.014
20	home_ownership	categorical	6.249	4	.181			
21	revol_util	continuous	.034	1	.855	1.000	.998	1.002

ويتبين عند هذه المرحلة أن المتغيرات المرشحة للتحليل متعدد المتغيرات كانت ١٦ (بعد حذف احدى صيغتي متغير مبلغ القرض المستمرة أو التصنيفية)، وأن المتغيرات التي غادرت التحليل كانت الأربعة الأخيرة.

٢-٦ التحليل متعدد المتغيرات

قُسمت مجموعة البيانات الكاملة (٤٢٤٢٣ حالة) إلى مجموعتين: مجموعة التدريب (٥٠%) التي تم منها تقدير النماذج، ومجموعة الاختبار (٥٠%) التي تم منها التحقق من مصداقية تلك النماذج. وتم تقدير الانحدار اللوجستي وشجرة القرارت.

*** الانحدار اللوجستي**

طُبق الانحدار اللوجستي على المتغيرات باتخاذ متغير حالة القرض كمتغير تابع (٠ سيء أو متعثر، ١ جيد أو غير متعثر) وباقي المتغيرات كمتغيرات مستقلة. واستُخدمت طريقة الانحدار المتدرج الأمامي المبنية على نسبة الإمكان بنقاط فصل[1] 0.3، ٠,٥، ٠,٧ على الترتيب. ويوضح جدول ٧ نتائج المقارنة بين تلك النماذج الثلاثة.وعلى الرغم أن القيم غير المعنوية لاختبار Hosmer andLemeshowتدل على قبول فرض العدم بملائمة نموذج الانحدار اللوجستي لوصف البيانات عند نقاط الفصل الثلاث، إلا أن أفضل النماذج كان عند نقطة القطع ٠,٣ (صاحب أعلى معدل دقة تصنيف للقروض السيئة).

جدول ٤: ملخص المقارنة بين نماذج الانحدار اللوجستي عند نقاط الفصل الثلاثة

Model	Cut value	Nagelkerke R^2	Hosmer and Lemeshow Test	Classification rates %		
				good	bad	Total
1	0.3	0.494	0.757	95.8	65.9	92.1
2	0.5	0.494	0.757	99.0	51.3	93.1
3	0.7	0.494	0.757	99.6	35.3	91.6

ويبين جدول ٥ أن أهم المتغيرات المؤثرة على الجدارة الائتمانية حسب النموذج الأول كانت ٧، هي: الدفعات الكلية، ومبلغ القرض، ودرجة التصنيف الائتماني، وأجل القرض، ومطابقة خصائص العميل للسياسة الائتمانية للبنك، والغرض من القرض، وحد التقاطع على الترتيب.

قُسمت مجموعة البيانات الكاملة (٤٢٤٢٣ حالة) إلى مجموعتين: مجموعة التدريب (٥٠%) التي تم منها تقدير النماذج، ومجموعة الاختبار (٥٠%) التي تم منها التحقق من مصداقية تلك النماذج. وتم تقدير الانحدار اللوجستي وشجرة القرارت.

*** الانحدار اللوجستي**

طُبق الانحدار اللوجستي على المتغيرات باتخاذ متغير حالة القرض كمتغير تابع (٠ سيء أو متعثر، ١ جيد أو غير متعثر) وباقي المتغيرات كمتغيرات مستقلة. واستُخدمت طريقة الانحدار المتدرج الأمامي المبنية على نسبة الإمكان بنقاط فصل[2] 0.3، ٠,٥، ٠,٧ على الترتيب. ويوضح جدول ٧ نتائج المقارنة بين تلك النماذج الثلاثة.

وعلى الرغم أن القيم غير المعنوية لاختبار Hosmer andLemeshow تدل على قبول فرض العدم بملائمة نموذج الانحدار اللوجستي لوصف البيانات عند نقاط الفصل الثلاث، إلا أن أفضل النماذج كان عند نقطة القطع ٠,٣ (صاحب أعلى معدل دقة تصنيف للقروض السيئة).

جدول ٤: ملخص المقارنة بين نماذج الانحدار اللوجستي عند نقاط الفصل الثلاثة

Model	Cut value	Nagelkerke R^2	Hosmer and Lemeshow Test	Classification rates %		
				good	bad	Total
1	0.3	0.494	0.757	95.8	65.9	92.1
2	0.5	0.494	0.757	99.0	51.3	93.1
3	0.7	0.494	0.757	99.6	35.3	91.6

ويبين جدول ٥ أن أهم المتغيرات المؤثرة على الجدارة الائتمانية حسب النموذج الأول كانت ٧، هي: الدفعات الكلية، ومبلغ القرض، ودرجة التصنيف الائتماني، وأجل القرض، ومطابقة خصائص العميل للسياسة الائتمانية للبنك، والغرض من القرض، وحد التقاطع على الترتيب.

: Variables in the Equation◦Table

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 6[a]	term(1)	1.686	.115	215.469	1	.000	5.397
	credit_policy(1)	.689	.124	30.658	1	.000	1.992
	grade			305.906	6	.000	
	grade(1)	2.699	.353	58.426	1	.000	14.866
	grade(2)	2.479	.244	103.401	1	.000	11.926
	grade(3)	2.038	.164	154.702	1	.000	7.673
	grade(4)	1.809	.126	205.726	1	.000	6.104
	grade(5)	1.352	.115	137.154	1	.000	3.864
	grade(6)	.781	.112	48.405	1	.000	2.184
	purpose			32.146	13	.002	
	purpose(1)	-.474-	.514	.848	1	.357	.623
	purpose(2)	-.667-	.519	1.653	1	.199	.513
	purpose(3)	-.809-	.530	2.332	1	.127	.445
	purpose(4)	-.143-	.533	.072	1	.788	.867
	purpose(5)	-.527-	.526	1.006	1	.316	.590
	purpose(6)	-.880-	.529	2.766	1	.096	.415
	purpose(7)	-.546-	.536	1.035	1	.309	.580
	purpose(8)	-.600-	.553	1.177	1	.278	.549
	purpose(9)	-1.011-	.560	3.255	1	.071	.364
	purpose(10)	-.735-	.557	1.743	1	.187	.479
	purpose(11)	-.140-	.562	.062	1	.803	.869
	purpose(12)	-.333-	.608	.301	1	.583	.717
	purpose(13)	-1.019-	.583	3.051	1	.081	.361
	loan_amnt	.001	.000	1089.983	1	.000	1.001
	total_pymnt	-.001-	.000	1249.633	1	.000	.999
	Constant	-2.868-	.532	29.063	1	.000	.057

a. Variable(s) entered on step 6: purpose.

[1] تُصنف الحالة في "نعم" إذا كان احتمالها المُتنبأ به باستخدام النموذج يزيد عن قيمة القطع المحددة.

*** شجرة القرارات**

أُجريت محاكاة الكمبيوتر لثلاث طرق مختلفة لشجرة القرارات (chi squared, entropy reduction, Gini reduction)، واختُبرت دقة تصنيفها للقروض الجيدة والسيئة.وحُسبت معدلات دقة التصنيف الصحيح على المستوى الكلى، وللقروض المسددة، وللقروض المتعثرة لثلاث نقاط فصل احتمالية (0.3, 0.5, 0.7). إذ قد يعتمد اختيار أفضل نموذج على قطع العتبة الاحتمالية التي يستخدمها البنك. ولأن الحدث المستهدف هو اكتشاف التخلف عن سداد القرض، فإن القطع عند الاحتمال 0.3 يعني أن تكلفة الخطأ (الناجمة عن منح القرض وهو لا ينبغي أن يُمنح) تكون أعلى ٣,٣ مرة(تحقق) من تكلفة رفضه عندما ينبغي أن يُمنح. وبمعنى آخر، فإن القطع عند ٠,٣ سوف يسمح للبنك بحذف معظم العملاء الذين يحتمل تعثرهم في سداد القرض ومنح الائتمان للعملاء الأكثر جدارة ائتمانية.كما يعني القطع ٠,٥ أن تكلفة الخطأ (الناجمة عن منح القرض وهو لا ينبغي أن يُمنح) تساوي تكلفة رفضه عندما ينبغي أن يُمنح. وأخيراً، فإن القطع ٠,٧ يعني أن تكلفة الخطأ (الناجمة عن منح القرض وهو لا ينبغي أن يُمنح) تكون أقل ٣,٣ مرة من تكلفة رفضه عندما ينبغي أن يُمنح. وبينما يناسب القطع عند ٠,٣ حالات القروض الصغيرة غير المؤمنة (التي لا تشترط أي ضمانات)، فإن القطع عند ٠,٥ و ٠,٧ يناسب القروض الأكبر(التي يُأخذ عليها ضمانات) مثل القرض بغرض شراء سيارةأو منزل.

وقد تبين عدم وجود اختلاف معنوي بين معدلات دقة التصنيف الصحيح للخوارزميات الثلاث عند نقطتي القطع ٠,٣ ، ٠,٥ سواء للقروض الجيدة (المسددة) أو السيئة (المتعثرة) أو على المستوى الكلي. وكانت خوارزمية CRT هي الأفضل في التصنيف الصحيح على المستوى الكلي لأن معدل دقة التصنيف لها بلغ ٩٢,٦%، بينما كانت خوارزمية CHAIDهي الأفضل في التصنيف الصحيح للقروض السيئة لأن معدل دقة التصنيف ٥٥,١%.

كما دُرست الأهمية النسبية للمتغيرات المستخدمة في بناء النماذج الثلاث لشجرة القرارت، وتبين أن النماذج الثلاث قد اتفقت على أن متغير "نسبة الدين إلى الدخل" كان أهم المتغيرات في التنبؤ بنتائج المتغير الهدف (حالة القرض)، حيث بلغت الأهمية النسبية للمتغير في النماذج الثلاث ١. كما اتفقت النماذج الثلاث أيضاً على أهمية ٣ متغيرات أخرى في التنبؤ بحالة القرض. وكانت هذه المتغيرات هي: عدد خطوط الائتمان المتأخرة (الأقساط المتأخرة)، والعمر بالشهور من أقدم خط ائتمان(فترة التأخر بالشهور)، وقيمة الممتلكات الحالية. وقد بلغت الأهمية النسبية المتوسطة للمتغيرات الثلاث: ٠,٣٥، ٠,٢٧، و ٠,٢٤، على التوالي.

وقد اقتصر عمق الشجرة فيها على ٣ مستويات. وبالرغم من أن معدلات دقة التصنيف للطرق الثلاثة كانت غير معنوية إحصائياً عند نقاط القطع الثلاثة، إلا أنه يبدو أن طريقتي chi squared and entropy reductionكانتا أفضل من طريقة Gini reduction لأنهما أنتجتا شجرتين أبسط من حيث عدد القواعد المستخدمة في عدد الانشقاقات. إذ كانت عدد القواعد في الطريقتين الأوليين ٩ لكل منهما، بالمقارنة بـ ١٤ للطريقة الثالثة.

وتؤكد خريطة ROC النتائجالسابقة بشأن قوة النماذج الثلاثة؛ حيث كلما ارتفع المنحني أكثر وكان على اليسار، كلما كان أداء النموذج أفضل.واستناداً للتحليل الشامل لمعدلات التصنيف، يُنصح بتطبيق طريقة chi squared لأنها تنتج شجرة القرارات بقواعد أبسط وأقل –وبالتالي يسهل تفسيرها أكثر– من الطريقتين الأخريين.

Table 6: Classification- tree1

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	2174	1769	55.1%
	1	610	25264	97.6%
	Overall Percentage	9.3%	90.7%	92.0%
Test	0	862	784	52.4%
	1	237	10723	97.8%
	Overall Percentage	8.7%	91.3%	91.9%

Growing Method: CHAID

Dependent Variable: credit

Table 7: Classification- tree2

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	2085	1801	53.7%
	1	396	25422	98.5%
	Overall Percentage	8.4%	91.6%	92.6%
Test	0	929	774	54.6%
	1	190	10826	98.3%
	Overall Percentage	8.8%	91.2%	92.4%

Growing Method: CRT

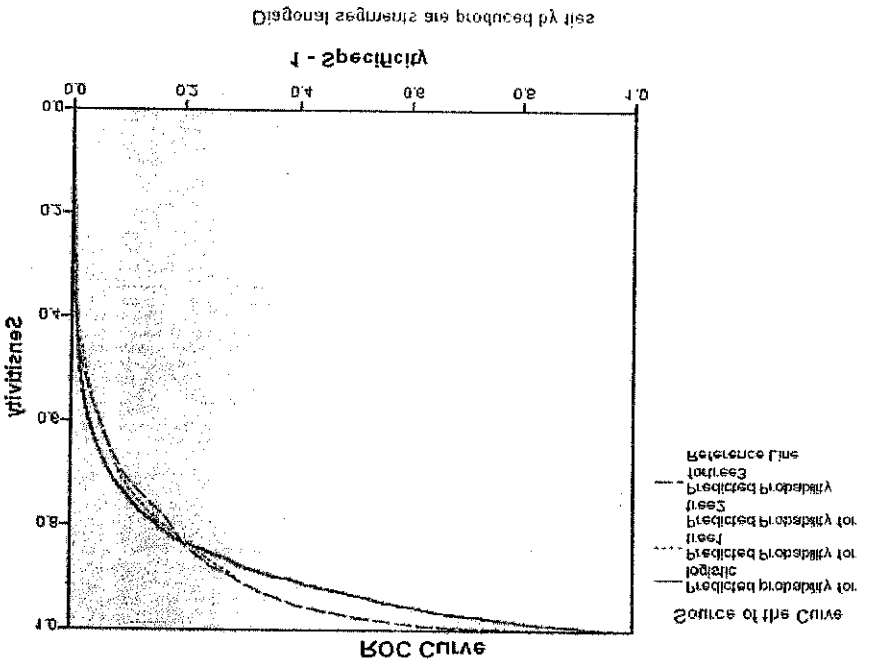
Dependent Variable: credit

Table 8: Classification- tree3

Sample	Observed	Predicted		
		0	1	Percent Correct
Training	0	1540	2303	40.1%
	1	336	25593	98.7%
	Overall Percentage	6.3%	93.7%	91.1%
Test	0	676	1070	38.7%
	1	159	10746	98.5%
	Overall Percentage	6.6%	93.4%	90.3%

Growing Method: QUEST

Dependent Variable: credit



شكل ٢: المقارنة بين دقة تصنيف الانحدار اللوجستي وأشجار القرار

٧) الخلاصة والمقترحات للبحوث المستقبلية

درس البحث البيانات التاريخية من القروض الاستهلاكية التي تصدرها إحدى المؤسسات المالية للأفراد الذين تعتبرهم عملاء مؤهلين للحصول على تلك القروض. ولما كان التصنيف الدقيق يفيد كل من الدائن (زيادة الربح أو تقليل الخسارة) والمدين (تجنب التعثر)، فقد استخدم البحث الانحدار اللوجستي وطرق استنتاج القاعدة (شجرةالقرارات) للتنبؤ بما إذا كان طالب القرض الجديد سيسدد قرضه أم سيتعثر فيه؛ وتفوق في ذلك الانحدار اللوجستي على أشجار القرارات. ويمكن تفسير هذه القواعد لكل من مديري الائتمان بالبنوك (الذين يحتاجون لفهمها قبل الموافقة على تنفيذها) وطالبي القروض كسبب لحرمانهم من القرض.

وينبغي إجراء مزيد من البحوث بشأن تحسين تدريب واختبار نماذج شجرة القرارات، إذ أن ضبط النماذج فضلا عن استخدام مجموعات البيانات المتوازنة وغير المتوازنة يساهم في تحسين أداء التصنيف. كما ينبغي مقارنة نتائج طرق شجرة القرارات بنتائج الطرق الأخرى للتنقيب في البيانات التي تناسب تقييم الائتمان (كالشبكات العصبية، والانحدار اللوجستي، والأنظمة الخبيرة،..) للوقوف على أفضلها في التصنيف الصحيح لحالتي القرض. وأخيراً، ينبغي على البنوك تحديد سياسة الإقراض الأكثر ربحية من منظور مخاطر الائتمان استنادا إلى الأرباح المتوقعة على القروض الجيدة، ومتوسط الخسائر على القروض السيئة، والتكاليف الثابتة والمتغيرة لعمليات الإقراض–أو تسمح للباحثين بحسابها بدلاً من اعتبار تلك البيانات موضوعات في غاية السرية.

١) مؤسسةالنقد العربي السعودي(٢٠١٣). التقرير السنوي، ١٩ فبراير ٢٠١٣.

2) Crook J, Hamilton R & Thomas L (1992). A comparison of discriminations under alternative definitions ofcredit default. In L C Thomas, J N Crook & D B Edelman (Eds.), Credit scoring and credit control, *Oxford: Oxford University Press, pp. 217–245.*

3) Dinh T and Kleimeier S (2007). "Acredit scoringmodel forVietnam's retail bankingmarket,"International Review of Financial Analysis, 16, pp. 471–495.

4) Fahrmeir L and Tutz G (1994).Multivariate Statistical Modelling Based on Generalised Linear Models.Springer, New York.

5) Feldman D, and Gross S (2005). "Mortgage Default: Classification Tree Analysis," Journal of Real Estate Finance and Economics, 30, pp. 369-396.

6) Giudici P (2003). Applied Data Mining: Statistical Methods for Business and Industry, John Wiley & Sons Inc, pp. 293-321.

7) Kamleitner B, Kirchler E (2007) , "Consumer credit use: a process model and literature review," Revue européenne de psychologie appliqué, 57, pp. 267–283.

8) Koh H, Tan W and Goh C(2006)."A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques,"International Journal of Business and Information, 1(1), pp. 96-118.

9) Lending Club (2013). Lending Club Statistics. https://www.lendingclub.com/info/download-data.action

10) Linoff and Berry (2011). Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management,3rd ed., Wiley, pp 169-199.

11) Sabzevari H, Soleymani M and Noorbakhsh E (2007). A comparison between statistical and Data Mining methods for credit scoring in case of limited available data. www.crc.man.ed.ac.uk/conference/archive/2007/papers/sabzevari-et-al.pdf

12) Thomas L(2000). "A survey of credit and behavioral scoring : forecasting financial risk of lending to consumers,"International Journal of Forecasting, 16(2), pp. 163-167.

13) Zurada J (2007). "Rule Induction Methods For Credit Scoring," Review of Business Information Systems – Second Quarter, 11(2), pp. 11-22.