Evaluation Statistique du Risque de l'Octroi du Crédit
Cas : de la Banque de l'Agriculture et du Développement Rural

Dr.Boudjenane Khaldia
Maitre de conférences classe « B »
Université Ibn Khaldoun –Tiaret-
dehbias60@gmail.com

Résumé :

Dans le présent travail nous essayerons de découvrir les méthodes quantitatives adoptées dans la classification et le traitement de demande des prêts bancaires offert aux entreprises et aux particuliers, parmi ces méthodes nous citons la régression logistique que nous allions appliquer sur 638 dossiers clients de la Banque de l'Agriculture et du Développement Rural – agence de Tiaret .

L'application de cette méthode nous a permit à déterminer les variables qui distinguent à mieux entre un bon et un mauvais client, et cela nous facilite de prendre la meilleur décision d'octroi du crédit.

Mots clés: la régression logistique, le crédit, les garanties, la décision, les méthodes quantitatives.

ملخص :

نهدف من خلال هذا المقال إلى كشف الغطاء عن الطرق الكمية المعتمدة في تصنيف المؤسسات و المتمثلة في تقنية الانحدار اللوجستي، و ذلك من خلال تطبيقها على 638 مؤسسة متعاملة مع بنك الفلاحة و التنمية الريفية - و كالة تيارت - بهدف تقدير خطر منح القروض بما يضمن اختيار أفضل لقرار الإقراض.

إن تطبيق مثل هذه الطرق الإحصائية مكننا من تحديد المتغيرات المميزة للتصنيف بين المؤسسات و الأفراد القادرين و غير القادرين على تسديد القرض، كما ساعدنا في اتخاذ القرار الرشيد.

الكلمات المفتاح : الانحدار اللوجستي ، القرض، الضمانات، القرار، الطرق الكمية.

INTRODUCTION :

Depuis la venue de l'informatique, l'ensemble des données ne cesse de croître. Les informations qui vont constituer la connaissance nécessaire à l'activité de l'entreprise sont disséminées. Ainsi, de nombreuses entreprises récoltent de plus en plus d'informations sur leurs clients et leurs comportements sans chercher à les exploiter. Pour cela, il est nécessaire de mettre en place des outils de traitement de données nouveaux pouvant tirer avantage de cette richesse enfouie. Les banques n'échappent pas à la règle, d'autant plus qu'elles sont régulièrement confrontées à des risques dont la gestion est reconnue aujourd'hui comme unes activité essentielle de ces dernières.

Le risque de crédit fut pendant longtemps un problème majeur pour les banques, car les mesures de contrôle qu'elles entreprennent pour faire face aux différents risques restent relativement peu développées. Donc les banques sont dans l'obligation de gérer leur risque de façon minutieuse de manière à prendre la bonne décision en un temps record.

Notre article s'inscrit dans le contexte de présenter l'automatisation de la règle de décision d'octroi de prêt d'une institution bancaire selon une méthodologie récente qui intègre plusieurs méthodes statistiques : le Data Mining.

Cela nous ramène à poser la problématique suivante :

Dans l'objectif d'aide à la décision des banques ***comment décider de l'attribution du crédit en utilisant les techniques statistiques ?***

Dans cet article, nous proposons une méthode de modélisation : la régression logistique, la méthode est appliquée respectivement sur les données recueillies, à partir de la banque B.A.D.R de Tiaret.

Pour réaliser cette étude, on mettra en pratique une démarche de fouille des données (Data Mining) .

I- PRESENTATION DU DATA MINING

Nous présentons dans cette première section, tout d'abord, la définition du Data Mining et nous exposerons, ensuite, la démarche du Data Mining.

La notion de data Mining est assez vague, imprégnée par des fausses idées et des confusions avec d'autre disciplines. Pour cela, une présentation précise s'avère indispensable.

I-1- Définition du Data Mining

Le Data Mining ou « fouille de données » est une étape de l'ECD (Extraction de Connaissances à partir de données) ou KDD (Knowledge Discovery in Databases).[1]

Le *Data Mining* est en fait un terme générique englobant toute une famille d'outils facilitant l'exploration et l'analyse des données contenues au sein d'une base décisionnelle de type Data Warehouse ou DataMart. Les techniques mises en action lors de l'utilisation de cet instrument d'analyse et de prospection sont particulièrement efficaces pour extraire des informations significatives depuis de grandes quantités de données.[2]

Le Data Mining est plus précisément l'ensemble des techniques descriptives et prédictives destinées à l'exploration et l'analyse de grandes bases de données, de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données.[3]

[1] - T. Benoit, **Projet de Data Mining**, CEREMADE, Unité Mixte de Recherche (UMR) n° 7534, CNRS et Université Paris-Dauphine, Paris, juin 2014, p 03.

[2] - G.Saporta, **Probabilités Analyse des Données & Statistiques**, 2eme Edition. Technip, Paris, 2006,P XXIIe.

[3] - J.X.Liu, **New Developments in Robotics Research**, Edition. Nova, New York, 2005, P 175.

Le Data Mining est une discipline dans laquelle on travaille en mode projet.[4] En bref, le Data Mining est l'art d'extraire des informations (ou plus précisément des connaissances) à partir des données.

I-2- La démarche du Data Mining

Nous décrivons dans ce qui suit les principales étapes suivies pour réaliser une étude de Data Mining.[5]

I-2-1- Définition des objectifs

Cette première étape consiste à étudier le problème afin de déterminer les objectifs à atteindre. Pour cela, il faut commencer par choisir le sujet, définir la population cible (les prospects, les clients, seulement les clients fidèles,...), préciser l'entité statistique étudiée (individu, famille, entreprise, ...), ainsi que les variables de l'étude et en particuliers le phénomène à prédire (client à risque et client sans risque,...). Il convient également de planifier le projet et de prévoir l'utilisation opérationnelle des informations extraites et des modèles produits.

I-2-2- Construction de la base d'analyse

Il s'agit dans cette étape de concevoir une base d'analyse qui servira à la construction des modèles à travers les données disponibles. Cette étape comprend :

- L'inventaire des données.
- L'organisation de la base d'analyse.

I-2-3- L'analyse préliminaire

Le fichier brut une fois constitué doit d'abord être « nettoyé » pour éliminer les erreurs et les incohérences. Il comporte alors en général un trop grand nombre de variables. Une exploration des liaisons entre chaque variable explicative et le critère à prédire permet en général d'éliminer les variables non pertinentes. On utilise alors des outils classiques : test du khi deux de liaison entre variables qualitatives, comparaison

[4] - R.Dominique, Data Mining, Université Paris Est la Vallée, Institut d'Electronique et d'Informatique ,paris,2017,p :04.

[5] - R. Lefebure, G. Venturi, Data Mining, Edition. Eyrolles, Paris, 2001, P 31-32

des pourcentages des modalités de la variable à prédire par chaque variable explicative.[6]

I-2-4- L'exploration des données

Après avoir construit la base d'analyse, on passe à la vérification des hypothèses prédéfinies à travers les techniques descriptives ou les techniques prédictives. Pour cela, il convient d'appliquer la procédure suivante :

- Etablir une pré-segmentation de la population étudiée ;
- Constructions des échantillons ;
- La mise en œuvre d'une ou plusieurs techniques de Data Mining.[7]

I-2-5- Validation et choix du modèle

Durant cette étape, les modèles qui ont été élaborées au cours des phases précédentes sont comparées et après l'interprétation des résultats obtenus, on retient le modèle le plus performant. Effectivement, avant tout déploiement, l'évaluation des modèles retenus doit se faire sur l'échantillon de validation.[8]

a- La procédure

La validation des modèles se fait selon deux optiques, l'inférence statistique et la validation prédictive.

En effet, Pour comparer les modèles de même nature, il existe des indicateurs statistiques spéciaux pour chaque méthode. Mais comme les indicateurs statistiques de deux modèles de natures différentes sont rarement comparables, on comparera les modèles à partir de leurs performances prédictives.

---

[6] -G. Saporta , **La Notation Statistique des E emprunteurs ou Scoring ,**Scribd, consulté le 15/11/2017, https://fr.scribd.com/document/56222221/Chap-It-Re-1.

[7] - R. El amin , Techniques de Data Mining pour la Gestion de la Relation Client dans les Banques, Faculté des Sciences Exactes & des Sciences de la Nature et de la Vie ,d'Département d'Informatique, Université Mohamed Kheider,Bisekra , juin 2014,p 14.

[8] - R. Lefebure, G. Venturi, **Op.cit.** ,P 33.

b- Les outils de validations prédictives

Pour évaluer la performance prédictive des modèles, on utiliser généralement le taux d'erreur de la matrice de confusion ou la méthode de validation croisée, toutefois il existe d'autre outils comme la courbe ROC, la courbe de Lift et leurs indices associés qui peuvent être utilisées dans le cas des modèles de classement en deux classes.

b-1- La matrice de confusion

La matrice de confusion est un tableau croisé où apparaissent en ligne la répartition de l'échantillon de validation par classe (valeurs réelles), et en colonne les résultats de classifieur (valeurs prédites). Le diagnostique de cette matrice représente les observations bien classées, (c'est-à-dire celles pour lesquelles la classe d'appartenance est la même), et les observations qui sont mal classées.[9]

Etant donné un classifieur et sa prédiction, 4 cas sont possibles :

- Vrai Positif (VP) : instance positive, classifiée positive ;
- Faux Positif (FV): instance négative, classifiée positive ;
- Vrai Négatif (VN): instance négative, classifiée négative ;
- Faux Négatif (FN): instance positive, classifiée négative.

On résume les résultats d'un classifieur sur des données de test dans une matrice de confusion sont montrées dans le tableau n°01 :[10]

Tableau n°01 : matrice de confusion.

Valeurs prédites Valeurs réels	Groupe1	Groupe2	Total
Groupe1	VP	VN	VP+VN
Groupe2	FP	FN	FP+FN
Total	VP+FP	VN+FN	**Total globale**

Source : J-P.Nacache, J.Confais, Statistique Explicative Appliquée, EditionTechnip, Paris, 2003, P 26.

[9] - SP. Roussel, F. Wacheux, **Management des Ressources Humaines**, Edition. de Boeck, Bruxelles, 2005, P 385.

[10] - D.Hosmer & Others, Applied **Logistic Regression**, 3rd Edition, Weley, New Jersey, 2016,p 145.

Le pourcentage des observations mal classées représente le taux d'erreur pouvant mesurer la qualité du modèle.

b-2- La courbe ROC

La courbe ROC est un outil d'évaluation et de comparaison des modèles à deux classes, indépendant des matrices de taux de mauvaise affectation. Il permet de savoir si un modèle M1 sera toujours meilleur que le modèle M2 quelle que soit la matrice de confusion.

La courbe ROC constitue en effet, un outil graphique qui permet de visualiser les performances prédictives d'un modèle, d'une telle façon où un seul coup d'œil doit permettre de voir le(s) modèle(s) susceptible(s) de nous intéresser.[11]

Un indicateur synthétique est associé à la courbe ROC, il s'agit de l'AUC (Aire Sous la Courbe). Il indique la probabilité d'un individu positif d'être classé devant un individu négatif. Il existe une valeur seuil, si l'on classe les individus au hasard, l'AUC sera égal à 0.5. Elle met en relation dans un graphique *les taux de faux positifs* (en abscisse) et *les taux de vrais positifs* (en ordonnée) ce qui veut dire que si on a sur la courbe le point (0.3, 0.9) cela signifie que le classifier a affecté 30% des faux positifs et 90% des vrais positifs.[12]

L'aire sous la courbe ROC est calculée par la formule :[13]

$$AUC = \sum_{i=1}^{\lfloor FP \rfloor - 1} \left(x_{i+1} - x_i\right) F_{ROC}\left(x_{i+1}\right)$$

Avec :

x_i : se sont les individus classer selon un score décroissant.

[11] - Galit Shmueli & Others **, Data Mining for Business Analytics: Concepts, Techniques, and Applications in R**, 1 st Edition ,weley , New Jersy, 2017,p 78.

[12] - A.D. Mezouar, **Recherche Ciblée de Documents sur le Web**, Thèse de doctorat de l'université Paris-Sud, Spécialité Informatique,2004,p 85.

[13] - Ricco Rakotomalala , **La Courbe Roc**, Tutoriels Tanagra ,15/11/2017, https://eric.univ-lyon2.fr/~ricco/cours/slides/roc_curve.pdf

$F_{ROC}(x_{i+1})$: Représente l'ordonnée de la courbe ROC au point x_{i+1}, telle que :

- $F_{ROC}(x_1) = 0 \qquad si \qquad x_1 \in FP$
- $F_{ROC}(x_1) = \frac{1}{\lfloor VP \rfloor} \qquad si \qquad x_1 \in VP$
- $\forall i \succ 0 \qquad F_{ROC}(x_{i-1}) + \frac{1}{\lfloor VP \rfloor} \qquad si \qquad x_1 \in VP$

Le degré de qualité de classement d'un classifieur est présenté dans le tableau n°02 :

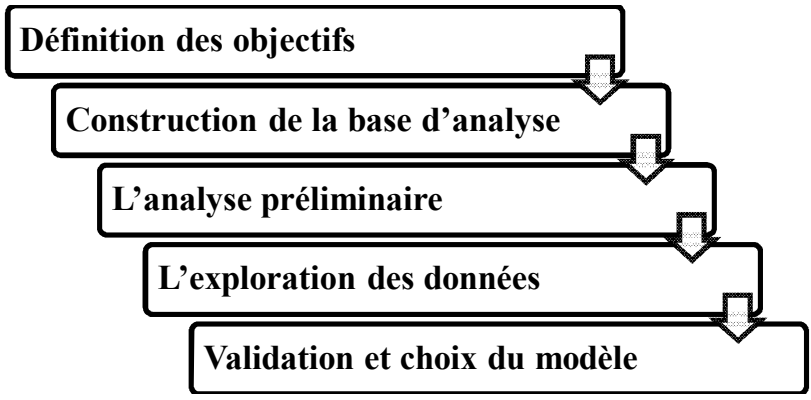
Tableau n°02 : Classification de la qualité d'un système à partir de l'aire sous sa courbe ROC.

Aire sous la courbe ROC	qualité
0.9 < AUC ≤ 1	Excellent
0.8 < AUC ≤ 0.9	Bonne
0.7 < AUC ≤ 0.8	moyenne
0.6 < AUC ≤ 0.7	médiocre
0 < AUC ≤ 0.6	mauvaise

Source : A.D. Mezouar,Loc.Cit.

Donc le Data Mining est une démarche simplifié et didactique en 5 temps majeurs.

Figure n°01 : Processus du data Mining.



Source :H.Jiawei & Autres, Data Mining : Concepts and Techniques, 3 eme Edition,2016,p :07.

II- Le prétraitement

La première phase du processus de Data Mining consiste principalement à préparer les données pour la modélisation. Cette tâche nécessite généralement un nettoyage et une transformation des variables pour supprimer toute sorte d'incohérence et d'atypisme, et pour assurer également l'homogénéité des variables.

Pour cela, nous commençons d'abord par présenter les données, puis nous essayons de les synthétiser et de révéler les liens existants entre la variable cible et les variables explicatives à travers l'analyse univariée et bivariée. Enfin, on procèdera à des nettoyages et transformations utiles sur les variables.

II-1- Présentation des données

Les données de cette application sont collectées à partir de tous les dossiers d'octroi de crédit trouvés au niveau de l'agence 541 de la banque BADR. La durée nécessaire était de un mois et demi, avec une fréquence d'une journée par semaine.

Le nombre de dossiers mis à notre disposition étaient de 635 dossiers collectés durant 12 ans .Dont 522 demandes de crédit acceptées et 113 demandes refusées par la banque. L'ensemble de ces crédits sont destinés au financement des projets d'investissement.

A partir de ces dossiers on a pu prélever 13 variables, une variable de type qualitative « à expliquer » (la variable cible), et 12 variables explicatives dont 8 sont qualitatives et 4 quantitatives. Les variables de l'étude associées avec le codage des variables qualitatives sont présentées dans le tableau n°03.

Tableau n°03 : Le codage des variables qualitatives.

Type des variables	Variables	Codage des variables
La variable qualitative à expliquer	Décision	(1) demande acceptée (0) demande refusée
Les variables explicatives qualitatives	Sexe d'emprunteur	(1) masculin (2) féminin
	Statut matrimonial	(1) célibataire (2) marié
	Niveau d'instruction	(1) faible «moyen et - » (2) moyen « lycée » (3) élevé «bac et + »
	Type de client	(1) particulier via l'ANSEJ ou CNAC (2) Entreprise via l'ANSEJ ou CNAC (3) particulier
	Zone d'activité	(1) urbaine (2) rurale (3) industrielle
	Type d'activité	(1) industrielle (2) service (3) artisanale (4) agricole (3) autre
	Type de projet	(1) création d'une entreprise ou micro entreprise (2) d'extension «augmentation des capacités de production» (3) renouvellement de l'outil de production Activité
	Rentabilité de projet	(1) faible (2) moyenne (3) forte

Source : établi a partir des données obtenus auprès de la banque BADR.

La rentabilité prévue d'un projet est apprécié par le comité de la banque à travers les tableaux de comptabilité.

a- La variable explicative quantitative discrète

- Année de la déclaration.

b- Les variables explicatives quantitatives continues

- Age d'emprunteur.

- Montant de crédit.
- La valeur des garanties matérielles.

Les garanties mis à la disposition de la banque se présentent sous les formes suivantes :

- **Hypothèque** (terrain, immeuble, magasin, écurie, ...) ;
- **Nantissement** (équipement, véhicule, assurance,...) ;
- **Cautionnement** (fond de garantie, tiers personne,...).

Les variables explicatives retenues peuvent être divisées en trois catégories :

- **Les variables qui décrivent l'emprunteur (**sexe, l'âge, statut matrimonial et le niveau d'instruction.)
- **Les variables liées au projet (**zone d'activité, type d'activité, type de projet, rentabilité de projet et le montant de crédit.)
- **Autre types de variables (**type de client, la valeur des garanties et l'année de déclaration.)

II-2- L'analyse bivariée

A l'aide des outils de l'analyse bivariée (test d'indépendance du khi deux et le rapport de corrélation entre variable quantitative et variable qualitative), on essaye de vérifier l'existence des liens entre la variable à expliquer « décision » et les variables explicatives.

En premier lieu, on établit le test d'indépendance du khi deux entre la variable à expliquer et les variables explicatives qualitatives ainsi que la variable quantitative discrète au seuil $\alpha = 0{,}05$, ensuite on calcule le rapport de corrélation entre variable quantitative et variable qualitative.

II-2-1- Test d'indépendance du khi deux

A travers le test d'indépendance du khi deux on va procéder au test des hypothèses suivantes :

- H_0 : la variable décision dépend de la variable explicative testé.
- H_1 : la variable décision est indépendante de la variable explicative testé.

Pour réaliser ce test, on est amené à construire un tableau de entre la variable décision et chaque variable explicative. Puis, on calcule la statistique de khi deux donnée sous H_0 par :

$$\chi^2 = \sum_{k=1}^{K}\sum_{l}^{L} \frac{n_{kl} - \frac{n_k \times n_l}{n}}{\frac{n_k \times n_l}{n}} \approx \chi^2_{1-\alpha}\left[(K-1)(L-1)\right]$$

Les résultats du test sont représentés dans le tableau suivant :

Tableau n°04 : Résultats du test d'indépendance de Khi deux.

Nom de la variable	P- Value	conclusion
Sexe	0,957	H_0 acceptée
Statut matrimonial	0,695	H_0 acceptée
Niveau d'instruction	0,000	H_0 rejetée
Type de client	0,019	H_0 rejetée
Zone d'activité	0,642	H_0 acceptée
Type d'activité	0,001	H_0 rejetée
Type de projet	0,011	H_0 rejetée
Rentabilité de projet	0,0001	H_0 rejetée
Année de déclaration	0,0001	H_0 rejetée

Source : résultats obtenus à partir du logiciel S.P.S.S.

Les résultats de test de khi deux montrent l'indépendance entre la variable décision et les variables sexe, statut matrimonial et zone d'activité. Par ailleurs, ils indiquent l'existence d'une dépendance entre cette variable d'intérêt et les variables niveau d'instruction, type de client, type d'activité, rentabilité et l'année dans laquelle la décision a été prise.

Dans première partie des résultats, on constate que la prise de décision d'octroi de crédit ne prend pas en compte le sexe de client, sa situation familiale et lieu d'activité dans lequel il va réaliser son projet. Cependant, la seconde partie des résultats, reflète l'existence d'un rôle très important des variables dépendantes précitées dans la prise de décision.

II-2-2-Le rapport de corrélation

Pour mesurer l'intensité de la relation entre la variable décision qui est qualitative et les variables explicatives quantitatives, on est amené à calculer le rapport de corrélation.

En effet, Le rapport de corrélation découle directement de la décomposition de la variance, il se définit dans le cadre générale pour une variable quantitative notée y et une variable qualitative notée x à K modalités de la manière suivante :

$$S^2{}_{y/x} = \frac{\sigma_E^2}{\sigma_y^2} telque 0 \le S^2{}_{y/x} \le 1$$

Avec :

$$\sigma_y^2 = \frac{1}{N}\sum_{h=1}^{K} n_h \left(\overline{y_h} - \overline{y}\right)^2 + \frac{1}{N}\sum_{i=1}^{K} n_h \sigma_i^2 = \sigma_E^2 + \sigma_R^2$$

σ_y^2 : c'est la variance totale de la variable quantitative y.

Le premier terme de la décomposition de σ_y^2 noté σ_E^2 est appelé variance expliquée (par X) ou variance inter classe ; le seconde terme noté σ_R^2 est appelé variance résiduelle ou variance intra classe. La règle d'interprétation est la suivante :

- $S_{y/x} = 0 \Rightarrow$ Absence de liaison entre x et y.
- $S_{y/x} = 1 \Rightarrow$ Liaison parfaite.

Les résultats du calcul de rapport de corrélation pour la variable qualitative à expliquer «décision » et les variables explicatives quantitatives sont représentés dans le tableau ci-dessous :

Tableau n°05 : Rapport de corrélation

variable	rapport de corrélation
Age d'emprunteur	0,0001
Montant de crédit	0,0964
valeurs des garanties	0,0089

Source : résultats obtenus à partir du logiciel S.P.S.S.

On remarque que le rapport de corrélation est proche de 0 pour les trois variables, ce qui implique l'absence de liaison entre les variables explicatives quantitatives et la variable à expliquer. Cela veut dire, que ces trois variables sont peu dispersées entre les deux groupes qu'à l'intérieur de chaque groupe, et donc elles expliquent faiblement la différence entre les deux groupes.

II-2-3-Nettoyage et transformation des données

Les variables explicatives de notre application prennent deux formes quantitatives ou qualitatives. Dans le but d'élaborer un modèle prédictif, on est amené à vérifier la pertinence de nos données en matière de prédiction, afin d'établir par la suite une transformation des variables quantitatives en qualitatives par un découpage en classes.

Les demandes de crédits sont généralement refusées à cause de deux facteurs, le premier est lié aux caractéristiques de l'emprunteur et à son projet, le second est lié seulement à la gestion et aux orientations de la banque.

Alors, pour construire un modèle prédictif qui permet de distinguer entre les clients à risque et les clients sans risque, on doit supprimer les emprunteurs refusés à cause des problèmes de la banque dont le nombre est de 31, ce qui veut dire que le nombre de clients intervenant dans la modélisation sera de 604 clients, 522 demandes acceptées et 82 demande refusées.

En outre, la variable année de déclaration ne peut pas être intégrée dans le modèle prédictif car elle se limite à un temps passé. Pour cela, elle doit être éliminée.

La transformation des variables quantitatives en qualitatives est très recommandée car elle permet de s'adapter facilement à certaines méthodes qui n'acceptent que des variables explicatives quantitatives. Toutefois, une telle transformation présente des problèmes liés aux choix de découpage:

- Ou bien on effectue, pour ne pas perdre trop d'information un découpage fin comportant beaucoup de classes de faibles amplitudes, mais au risque d'avoir des résultats instables et peu reproductibles du fait de classe d'effectif très faible.
- Ou bien on effectue un découpage grossier en un nombre de tranches relativement restreint, ces tranches étant de grande amplitude auquel cas on perd beaucoup d'information.

Pour éviter ces écueils, on opte pour un découpage proposé par défaut dans le logiciel de SPSS, qui consiste à découper les variables quantitatives en quatre classes par l'intermédiaire des quartiles. Cette procédure permet, en effet, d'obtenir un nombre modéré de classes, dont le nombre d'effectif et assez important (25% de l'ensemble des observations).

Après le découpage des variables quantitatives : âge d'emprunteur, montant de crédit et valeur des garanties, on a obtenu les variables qualitatives ordinales suivantes : âge d'emprunteur montant de crédit, valeur des garanties.

Par ces opérations, le nombre d'observation se restreint à 604 clients dont 522 demandes acceptées et 82 demandes refusées. Le nombre de variable se diminue également à 12 variables qualitatives, 1 variable cible « décision » et 11 variables explicatives.

III- L'exploration des données

L'exploration des données constitue le cœur de notre application. En effet, cette étape va nous permettre de construire le modèle prédictif qui répond aux objectifs prédéfinis.

On procède pour cela, à partitionner la population en échantillon d'apprentissage et en échantillon de validation. Après, on met en œuvre la méthode de modélisation : la régression logistique.

III- 1- Construction des échantillons

L'échantillonnage est une étape incontournable dans l'application notamment pour la prédiction ou le classement, plusieurs techniques sont possibles mais comme la taille de nos données (604 individus) est relativement petite, de plus les logiciels disponibles ne permettent pas de réaliser une validation croisée, on procède à un partitionnement de la population en deux échantillons ; un échantillon d'apprentissage de 75% de la population (453 individus) et un échantillon de validation de 25% de la population (151 individus).

Pour réaliser le partitionnement, on a utilisé le sondage stratifié à allocation proportionnelle afin de garder les mêmes proportions d'individus pour chacune des classes à prédire, cela veut dire qu'on répartie la population en deux sous ensembles (strates) par rapport à la variable cible «décision». Cette répartition nous a donné deux strates de taille 522 et 82 respectivement.

On tire aléatoirement de chaque strate 75% d'individus qui vont former l'échantillon d'apprentissage, et 25% d'individus qui formeront l'échantillon de validation.

- **L'échantillon d'apprentissage :**

Soient :

N_{app} : la taille d'échantillon d'apprentissage.

$Nap\text{-}_{accep}$: la taille d'échantillon d'apprentissage tiré de la strate des demandes acceptées.

$Nap\text{-}_{ref}$: la taille d'échantillon d'apprentissage tiré de la seconde strate.

On a donc, $Nap\text{-}_{accep} = 0{,}75 \times 522 = 391$ individus

$Nap\text{-}_{ref} = 0{,}75 \times 82 = 62$ individus

D'où: $Napp = N_{ap\text{-}accep} + N_{ap\text{-}ref} = 391 + 62 = 453$ individus.

- **L'échantillon de validation :**

Soient :

N_{val}: la taille d'échantillon d'apprentissage.

$N_{val\text{-}accep}$: la taille d'échantillon d'apprentissage tiré de la strate des demandes acceptées.

$N_{val\text{-}ref}$: la taille d'échantillon d'apprentissage tiré de la seconde strate.

On a donc, $N_{val\text{-}accep} = 0{,}25 \times 522 = 131$ individus

$N_{ap\text{-}ref} = 0{,}25 \times 82 = 20$ individus

D'où: $N_{val} = N_{val\text{-}accep} + N_{val\text{-}ref} = 131 + 20 = 151$ individus.

La répartition de la population est récapitulée dans le tableau suivant :

Tableau n°06 : Taille des échantillons.

Echantillon	Proportion de la population	Taille	Décision	Taille
Apprentissage	0,75	453	Demande acceptée	391
			Demande refusée	62
Validation	0,25	151	Demande acceptée	131
			Demande refusée	20

Source : résultats obtenus à partir du logiciel XL Stat.

III - 2- La modélisation

L'élaboration du modèle sélectionné est réalisée à partir de l'échantillon d'apprentissage, suivi de la validation de ce modèle sur la base des données de l'échantillon de validation.

L'échantillon d'apprentissage est constitué de 453 individus, décrits par 12 variables qualitatives dont une variable est la variable cible « décision » et les autres sont des variables explicatives. En revanche, l'échantillon de validation est constitué de 151 individus décrits également par les 12 variables. La méthode qu'on va mettre en œuvre est *la régression logistique*.

III- 2- 1- La régression logistique

Les données statistiques disponibles dans la base d'analyse sont relatives à des caractères qualitatives (Décision, type d'activité, sexe, ...), or les méthodes d'inférence

traditionnelle ne permettent pas de modéliser et d'étudier ces caractères, pour cela, on pense à l'utilisation de la régression logistique qui permet de modéliser ces caractères en tenant compte de l'absence de la continuité ou de l'ordre naturelle entre les modalités que peut prendre le caractère qualitatif.

En effet, la régression logistique est destinée à la modélisation d'une variable endogène qualitative expliquée par un ensemble de variables exogène quantitatives.

Toutefois, les variables explicatives qualitatives peuvent être utilisées après la transformation en forme disjonctive en supprimant une modalité pour chaque variable qualitative.

Dans cette application, on s'appuie essentiellement sur les résultats de logiciel XL Stat.

a- Construction du modèle

Cette application consiste à expliquer la variable dichotomique « décision » a deux modalités « demande acceptée et demande refusée » par 25 modalités provenant de 11 variables qualitatives.

La condition d'application du modèle Logit sur des variables explicatives qualitatives est bien vérifiée où on a supprimé une modalité de chaque variable afin d'assurer l'existence de la matrice de variance covariance.

On a :

- **Le modèle Logit**

qui s'écrit sous la forme : $P_i = F(x_i\beta) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} = \frac{1}{1+e^{-x_i\beta}}$

Avec P_i est la probabilité associée à une demande acceptée pour l'individu i.

- **Test de Wald :** $\frac{\beta_j^2}{S^2(\hat{\beta}_j)} \to \chi^2 \qquad sous \qquad H_0$

On rejette H_0 si : $W_j = \frac{\hat{\beta}_j^2}{S^2(\hat{\beta}_j)} \succ \chi_\alpha^2 \qquad \Leftrightarrow \Pr\underbrace{[\chi_{(1)}^2 \succ W_j}_{P\_Value}] \prec \alpha$

$$\Leftrightarrow \frac{|\hat{\beta}_j|}{S(\hat{\beta}_j)} \succ u_{\frac{\alpha}{2}}$$

Les résultats de l'estimation sont récapitulés dans le tableau suivant :

Tableau n°07: Estimation des paramètres du modèle Logit.

Source	Valeur	Ecart-type	Khi² de Wald	ddl	Pr > Khi²
Constante	8,647	126767,41	0,000	1	1,000
Masculin	-2,636	1,418	3,457	1	**0,063**
Célibataire	2,399	0,877	7,489	1	**0,006**
nv_faible	-0,274	1,221	0,050	1	0,822
nv_moyen	-0,584	1,126	0,269	1	0,604
particulier_ansej-cnac	-15,68	9525,397	0,000	1	0,999
entreprise_ansej-cnac	24,472	130886,209	0,000	1	1,000
Zone urbaine	-15,289	19065,886	0,000	1	0,999
Zone rurale	-11,45	19065,886	0,000	1	1,000
act_industrielle	-1,890	1,938	0,952	1	0,329
act_service	-0,015	1,843	0,000	1	0,994
act_artisanale	-0,607	1,959	0,096	1	0,757
act_agricole	-2,469	2,108	1,371	1	0,242
prj_création d'entreprise	27,962	124962,942	0,000	1	1,000
prj_extension	-11,19	37746,238	0,000	1	1,000
rentab_faible	-47,707	14196,703	0,000	1	0,997
rentab_moyenne	-4,304	1,138	14,294	1	**0,000**
âge-[19, 27[	-1,032	1,077	0,919	1	0,338
âge-[27, 31[	1,179	1,216	0,940	1	0,332
âge-[31, 35[	0,089	0,824	0,012	1	0,914
montant-[139285, 1054011[	1,667	0,966	2,978	1	**0,084**
montant-[1054011, 1415473[	18,985	2398,706	0,000	1	0,994
montant-[1415473, 2093714[	-0,999	0,939	1,131	1	0,288
garanties-[0, 1500898[	2,528	0,894	8,000	1	**0,005**
garanties-[1500898, 2183093[	20,887	2498,659	0,000	1	0,993
garanties-[2183093, 3694068[	5,859	1,354	18,714	1	< **0,0001**

Source : résultats obtenus à partir du logiciel XL Stat.

Le test de Wald de significativité des paramètres indique que 6 paramètres sont significativement différents de 0 au seuil $\alpha = 0.1$, ce qui veut dire que les modalités associées sont discriminantes, et apporte une certaine information sur la variable cible.

Les 6 modalités discriminantes proviennent de 5 variables explicatives qui sont : sexe masculin, statut matrimonial célibataire, moyenne rentabilité de projet, le montant de crédit entre [139285, 1054011[et les garanties à valeur dans l'intervalle [1500898, 2183093[et [2183093, 3694068[.

b- Critère de spécification du modèle

Les principaux indicateurs[14] offerts par XL Stat sont :

Tableau n°08 : Critère de la spécification du modèle Logit.

Critère	Valeur
R^2(McFadden)	0,748
R^2 (Cox and Snell)	0,449
R^2(Nagelkerke)	0,817
AIC	14,313
SBC	250,326

Source : résultats obtenus à partir du logiciel S.P.S.S.

Ces critères montrent dans l'ensemble, que le modèle obtenu est acceptable et plus efficace qu'un modèle obtenu par la méthode sélection des variables pas à pas.

[14] - R. Rakotomalala, **Pratique de la Régression Logistique**, Université Lumière Lyon 2, p 35,consulte le 06/11/2017,
https://eric.univlyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf

c- les signes des paramètres

On représente dans le tableau suivant les signes des paramètres significatifs du modèle :

Tableau n°09 : Singes des paramètres du modèle Logit.

Modalité	Signe de paramètre
Masculin	-
Célibataire	+
Rentabilité moyenne	-
Montant : [139285, 1054011[	+
Garanties : [0, 1500898[	+
Garanties :[2183093, 3694068[	+

Source : résultats obtenus à partir du logiciel XL Stat.

On remarque que deux modalités : sexe masculin et la rentabilité moyenne de projet influent négativement sur la probabilité associée à une demande acceptée. Par contre, les modalités : célibataire, montant de crédit entre [139285, 1054011[et les garanties entre [0, 1500898[et [2183093, 3694068[influent positivement. Cela veut dire que, les emprunteurs à sexe masculin qui veulent réaliser des projets à rentabilité moyenne sont désavantagés. En revanche, les célibataires qui demandent un crédit à faible montant (*inférieur à 1 millions DA*) en proposant une garantie importante ont plus de chance de se voir accorder un crédit.

d- Le rapport de chance «Odds ratio»

Le calcule des effets marginaux permettent de mesurer la part de déviation des modalités par rapport à la situation de référence.

Dans le tableau suivant on donne le rapport de chance (Odds ratio) des modalités des variables significatives en mettant en dernier la modalité de référence.

$$Odds_p = \frac{P}{P-1}$$

Tableau n°10 : Rapport de chance.

	Variable	Odds ratio	IC Borne inf. (95%)	IC Borne sup. (95%)
Sexe	masculin	**0,072**	0,004	1,154
	féminin	-	-	-
Statut matrimonial	Célibataire	**11,010**	1,975	61,361
	marie	-	-	-
Rentabilité	moyenne	**0,014**	0,001	0,126
	forte	-	-	-
Montant de crédit	[139285, 1054011[	**5,297**	0,797	35,190
	[2093714, 20147331]	-	-	-
Valeurs des garanties	[0, 1500898[	**12,522**	2,173	72,165
	[2183093, 3694068[	**350,445**	24,645	4983,334
	[3694068, 9895016165]	-	-	-

Source : résultats obtenus à partir du logiciel XL Stat.

A partir des signes des paramètres significatifs, on constate que deux modalités influent à la baisse la probabilité pour obtenir un crédit, donc une faible chance d'acquérir un crédit.

D'après les résultats, les hommes ont 14 fois moins de chance que les femmes d'obtenir un crédit. Cela peut se justifier par les orientations de l'Etat qui visent à encourager l'activité féminine. De plus, un projet à moyenne rentabilité a 71 fois moins de chance d'être financé par la banque. Ce résultat, exprime la stratégie prudentielle suivie par la banque lors de traitement des dossiers pour accorder un prêt.

En revanche, 4 modalités provenant de 3 variables augmentent les chances de l'emprunteur pour obtenir un crédit. Effectivement, les emprunteurs célibataires ont 61 fois plus de chances que les mariés d'obtenir un crédit. Ce résultat revient généralement à l'encouragement de l'emploi des jeunes visé par l'Etat. De plus, on remarque que les emprunteurs qui sollicitent un montant de crédit dans les environs 1 millions DA possèdent 5 fois plus de chance d'obtenir un crédit que les emprunteurs

qui demandent plus de 2,13 millions DA, ce qui est évident car l'augmentation du montant signifie une augmentation des risques de non remboursement.

Enfin, mettre à la disposition de la banque une somme de garantie entre 0 et 1,5 millions DA ou entre et 2,2 et 3,7 millions DA donne à l'emprunteur 12 fois plus de chance pour obtenir un crédit que ceux qui mettent plus de 3,5 million DA. Ce résultat exprime, que, le fait de mettre un grosse somme de garanties implique forcement une demande de crédit assez conséquente. Par contre, le résultat précédent montre clairement que la banque défavorise les crédits à montant élevé.

e- Qualité de la représentation

La qualité de la représentation va nous permettre de quantifier le degré de performance du modèle Logit dans le reclassement des individus de l'échantillon d'apprentissage dans leurs groupes respectifs. Pour cela, on analyse la matrice de confusion qui regroupes les individus bien classés et ceux mal classés :

Tableau n°11 : Matrice de confusion sur échantillon d'apprentissage pour le modèle Logit.

de \ Vers	demande acceptée	demande refusée	Total	% correct
demande acceptée	382	9	391	97,70%
demande refusée	14	48	62	77,42%
Total	396	57	453	94,92%

Source : résultats obtenus à partir du logiciel XL Stat.

La matrice de confusion montre que 94.92% des emprunteurs de l'échantillon d'apprentissage sont bien reclassés, dont 97.70% des emprunteurs admis et 77.42% des emprunteurs refusés ont été affectées correctement à leurs classes.

Le taux d'erreur de reclassement est de 5.08%, ce qui veut dire que la probabilité d'effectuer une mauvaise prédiction à l'aide de ce modèle est de 5.08%.

f- Validation prédictive

Pour évaluer la performance prédictive du modèle obtenu, on utilise l'échantillon de validation pour calculer les indicateurs suivants :

➢ **la matrice de confusion pour l'échantillon de validation**

Tableau n°12 : matrice de confusion sur échantillon de validation pour le modèle Logit.

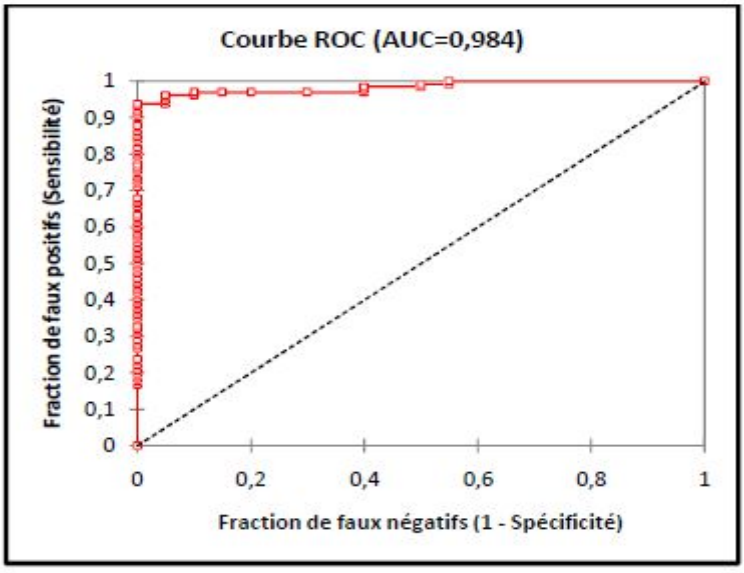
de \ Vers	demande acceptée	demande refusée	Total	correct%
demande acceptée	126	5	131	96,18%
demande refusée	1	19	20	95,00%
Total	127	24	151	96,03%

Source : résultats obtenus à partir du logiciel XL Stat.

La matrice de confusion montre que 96,07% des demandes de l'échantillon de validation ont été bien reclassées, dont 96,18% des demandes acceptées et 95,00% des demandes refusées ont été affectées avec succès. Alors, le taux d'erreur de reclassement est égal à 3,97%.

➢ **la courbe ROC et l'aire sous la courbe ROC**

Figure n°02 : la courbe ROC et l'aire sous la courbe ROC



Source : résultats obtenus à partir du logiciel XL Stat.

On remarque que la courbe ROC se situe au dessus de la diagonale avec une aire égale à 0.984 très proche de 1, alors la qualité de classification est excellente.

L'application de la régression logistique a permis de construire un modèle prédictif efficace (*taux d'erreur sur l'échantillon de validation est de 3.97%, et AIC =0.984*).

Parmi les 25 modalités initiales, 6 modalités sont significatives et interviennent dans la prise de décision en augmentant ou en diminuant les chances de l'emprunteur pour obtenir un crédit.

Les deux modalités sexe féminin et statut matrimoniale célibataire augmentent les chances du client à obtenir un crédit. Ce qui prouve que les orientations sociopolitique de l'Etat qui vise à soutenir l'activité féminine et en général l'emploi ;des jeunes, influent sur la décision d'octroi de crédit.

Les projets à rentabilité moyenne possèdent moins de chance à être acceptés que les projets à forte rentabilité. De plus, un crédit à montant relativement faible (inférieur à 105 millions) possède plus de chance à être accepté que des crédits à montant plus élevé. Donc, la qualité d'un projet proportionnellement à son montant joue un rôle très important dans la prise de décision.

En dernier, les crédits couverts par une grande valeur de garanties, ne sont pas toujours acceptés, et cela revient principalement aux risques de non remboursement et aux orientations sociopolitiques de l'Etat.

CONCLUSION :

En conclusion, notre étude s'est portée sur la du Data Mining appliquée sur des données bancaires en vue de construire un modèle prédictif robuste permettant de mieux distinguer entre un bon et un mauvais client.

L'opération d'octroi de crédit nécessite effectivement de collecter des informations sur l'emprunteur afin d'évaluer sa solvabilité et de décider s'il s'agit d'un bon ou d'un mauvais payeur.

Dans un second temps nous avons consacré notre étude à la mise en œuvre du processus de data Mining sur les données des dossiers de prêts, un prétraitement des données fut nécessaire en vue de décrire et de préparer les données à la phase de modélisation. Par la suite, on s'est servi du sondage stratifié pour scinder l'ensemble

des observations en deux échantillons : un échantillon d'apprentissage, et un échantillon de validation. Enfin, la méthode de modélisation Logit.

D'après le traitement et l'analyse des données, on a pu tirer les conclusions suivantes:

- La méthode de la régression logistique est la meilleure méthode prédictif, dont la probabilité de commettre une erreur de prévision est égale à **3.97%** .
- Les principaux caractères influant sur la décision d'octroi de crédit sont: le ***sexe*** *(masculin)*, le ***statut matrimonial*** *(célibataire)*, la ***rentabilité de projet*** *(moyenne)*, le ***montant de crédit*** (inférieur à 1 million DA) et enfin les ***garanties*** (à valeur inférieur à 1,5 millions DA ou entre 2,2 et 3,7 millions de DA).
- Les orientations sociopolitiques de l'Etat jouent un rôle considérable dans la prise de décision, cela se justifie, d'une part, par l'évolution du nombre de demandes de crédits par année et d'une autre part par la volonté de l'Etat à encourager les jeunes et surtout les célibataires.
- La qualité du projet d'investissement constitue un facteur primordial dans la prise de décision, de telle sorte que les projets à rentabilité faible sont directement refusés et les projets à rentabilité forte ou moyenne sont généralement acceptés ;
- La soumission au règlement de l'Etat favorise les emprunteurs à acquérir le crédit lorsque le projet présente une rentabilité acceptable.
- Le montant du crédit joue un rôle crucial dans la décision finale, généralement la banque défavorise les crédits à montant élevé même si l'emprunteur présente une grande valeur de garantie. Par conséquent, une grande somme de garantie ne signifie pas que le crédit sera accepté.

REFERENCES :

1. A.D. Mezouar, Recherche Ciblée de Documents sur le Web, Thèse de doctorat de l'université Paris-Sud, Spécialité,2004.
2. D.Hosmer & Others, Applied Logistic Regression, 3rd Edition, Weley, New Jersey, 2016.
3. G.Saporta, Probabilités Analyse des Données & Statistiques, 2eme Edition. Technip, Paris, 2006.
4. Galit Shmueli & Others , Data Mining for Business Analytics: Concepts, Techniques, and Applications in R, 1 st Edition ,weley , New Jersy, 2017.
5. H.Jiawei & Autres, Data Mining : Concepts and Techniques, 3 eme Edition,2016.
6. J.X.Liu, New Developments in Robotics Research, Edition. Nova, New York, 2005.
7. J-P.Nacache, J.Confais, Statistique Explicative Appliquée, EditionTechnip, Paris, 2003.
8. R. El amin , Techniques de Data Mining pour la Gestion de la Relation Client dans les Banques, Faculté des Sciences Exactes & des Sciences de la Nature et de la Vie ,d'Département d'Informatique, Université Mohamed Kheider,Biskra , juin 2014.
9. R. Lefebure, G. Venturi, Data Mining, Edition. Eyrolles, Paris, 2001.
10. R.Dominique, Data Mining, Université Paris Est la Vallée, Institut d'Electronique et d'Informatique, paris,2017.
11. SP. Roussel, F. Wacheux, Management des Ressources Humaines, Edition. de Boeck, Bruxelles, 2005.
12. T. Benoit , Projet de Data Mining, CEREMADE , Unité Mixte de Recherche (UMR) n° 7534, CNRS et Université Paris-Dauphine, Paris, juin 2014.

13. G. Saporta , La Notation Statistique des E emprunteurs ou Scoring Scribd, consulté le 15/11/2017, https://fr.scribd.com/document/56222221/Chap-It-Re-1.
14. R. Rakotomalala, Pratique de la Régression Logistique, Université Lumière Lyon 2, p 35,consulte le 06/11/2017, https://eric.univlyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf
15. Ricco Rakotomalala , La Courbe Roc , Tutoriels Tanagra ,15/11/2017, https://eric.univ-lyon2.fr/~ricco/cours/slides/roc_curve.pdf