Credit Scoring with Machine Learning
Approche de modélisation et intégration au processus d'octroi de prêts

Table des matières

1. Introduction : Use Case

Le projet a pour cadre la décision d'octroyer ou non un prêt bancaire, dont l'Use Case primaire peut être décrit comme suit :

Figure 1 : Use Case primaire

Un client ***« Applicant »*** soumet une demande de prêt à l'organisme : la demande est analysée et donne lieu au refus ou à l'accord de prêt.

L'organisme collecte alors le motif du **refus** et, dans le cas de prêts **accordés**, l'ensemble des données de remboursement.

L'objectif du projet est double :

1. Construire l'approche Machine Learning permettant de prédire, à partir des données collectées, une future difficulté de remboursement,
2. Intégrer cette approche dans le processus d'octroi de prêt de l'organisme par la fourniture d'un tableau de bord interactif.

Pour cela, commençons par détailler notre use case.

➔ ***NB. au fil du document, les perspectives sont identifiées avec cette syntaxe.***

2. Use Case détaillé

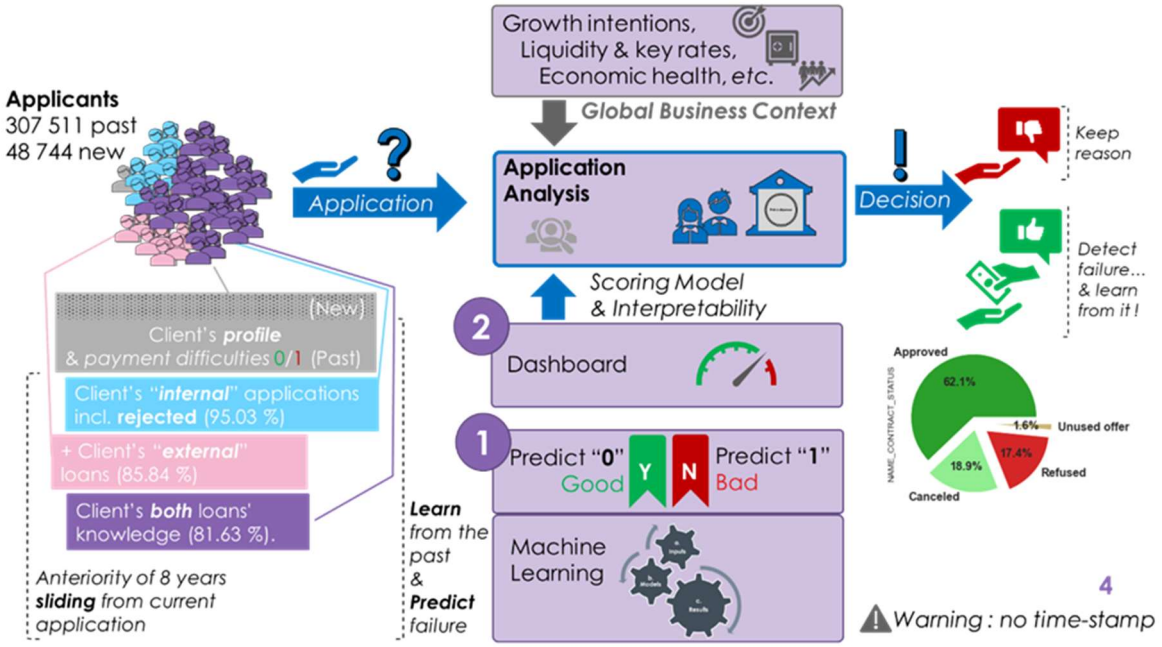


Figure 2 : Use Case détaillé

En accumulant au fil du temps le suivi des remboursements des prêts, l'organisme a construit un indicateur cible identifiant une difficulté de remboursement. Cette cible ***« TARGET »*** est connue pour 307 511 dossiers et consiste en une information binaire qui vaut 1 que lorsqu'un client a connu un défaut de payement ***« failure »*** durant plus de X jours pendant les Y premiers remboursements.

L'organisme dispose donc des caractéristiques du prêt demandé, des données de profil du client ainsi que **8 années** d'antériorité (glissantes) des emprunts internes ou externes contractés par le client.

Nous reviendrons ultérieurement sur l'analyse des données disponibles.

Les 2 objectifs du projet sont illustrés ci-dessus 1. ***Learn & Predict***, 2. ***Dashboard***.

Ils consistent à fournir et étayer pour chaque dossier parmi les 48 744 nouvelles demandes, une **prédiction de la cible**, **interprétable** comme un risque futur de défaut de paiement par le client.

On note que **17,4 % des demandes sont refusées**, et 62,1 % des demandes accordées sont effectivement mise en place. Cela ouvre un espace entre la « sévérisation » des refus et le risque de défaut de paiement réellement observé : plus de 2 fois plus de dossiers sont considérés porter un risque suffisamment élevé pour refuser la demande.

Le projet s'inscrit dans un contexte élargi des KPIs de l'organisme comme les stratégies de croissance, ainsi que des paramètres externes tels que les taux directeurs ou la santé économique.

Ces éléments de contexte sont inconnus et eux-mêmes fluctuants au cours du temps (et nous agissons en temps relatif), nous pouvons au mieux considérer qu'il existe des facteurs externes qui pilotent l'activité d'analyse/décision.

➔ ***Nous proposerons une base d'interprétation ouverte qui évoluera selon les objectifs à préciser.***

3. Données cible et sources

La cibe est l'information de classification binaire avec Class 1 : risk de defaut de paiement.

Les taux de défauts (failure rates) concernent environ 10% des cas, avec une prévalence des prêts classiques dits *'cash loans'* sur les prêts dits *'revolving loans'* ou renouvelable : lorsque l'organisme met à disposition une somme d'argent réutilisable au fur et à mesure du remboursement, pour effectuer des achats non prédéfinis.

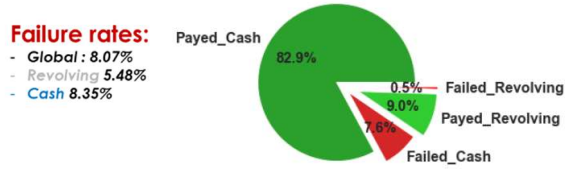


Figure 3 : taux de défauts de paiements

Le risque lié à un défaut de remboursement semble moindre dans le cas de prêts renouvelables, en effet, ces derniers engagent des moindres sommes, et offrent une possibilité d'interruption permettant de limiter les dommages financiers.

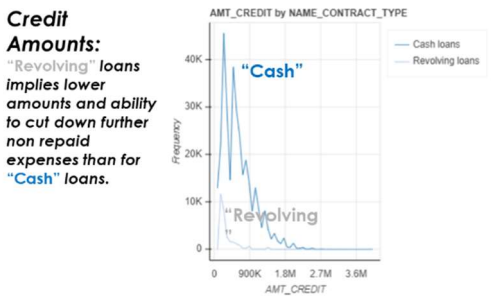


Figure 4 : montants des crédits

Dans notre étude, nous allons néanmoins considérer que le couple {X, Y}, qui traduit les données de paiement en cible binaire, a été décliné de façon à équilibrer pour les 2 types de prêts cette notion de risque associé à un défaut (target = 1).

➔ ***Une perspective serait toutefois d'étudier de façon distincte ces 2 cas de figures.***

Nous disposons de 7 jeux de **données sources** qui détaillent le « profil » du client et la description du prêt qu'il sollicite. En complément, comme illustré figure 5, l'antériorité client est disponible et atteint en cumul de sources externes et internes pour plus **80 % des dossiers**.

A cette étape du projet, il est demandé de réemployer un kernel Kaggle. Si cela ouvre un gain de temps substantiel pour avancer dans le projet, cela peut néanmoins conduire à handicaper la bonne marche du projet.

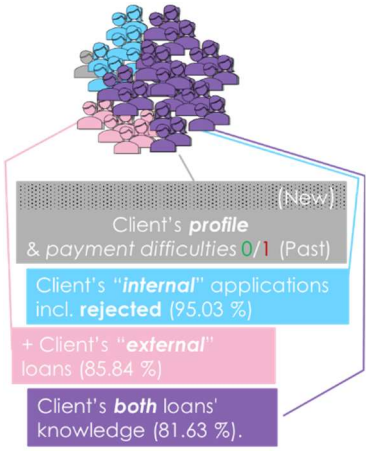


Figure 5 : données de profil et d'antériorité client

En particulier, tandis que les compétiteurs se focalisent sur l'obtention du meilleur score par tous les moyens, notre objectif est de construire une démarche **interprétable** et même nous efforcer de rendre **transparente** l'exploitation du Machine Learning.

En ce sens, le Feature Engineering et l'étape de Feature Selection font parties intégrantes de l'approche de modélisation et nous avons engagé une approche **Top-Down** pour l'exploration des données combinée à une approche **Bottom-Up** de Feature Engineering (manuel et automatique), pour aboutir à l'agrégation des données.

Les problématiques soulevées lors de cette phase préliminaire sont nombreuses et didactiques, sur les façons de préserver l'intégrité des features, de remédier au déséquilibre des classes ou encore d'imputer les valeurs manquantes de façon adéquate. Le tout étant **interdépendant**.

Une étape indépendante des modèles (dite model-agnostique) a également été menée à l'aide du package python Boruta.

➔ ***Une perspective, au-delà des pistes en réponse aux problématiques soulevées, serait toutefois d'itérer une fois le cadre de modélisation sélectionné (interdépendance).***

4. Panel des classificateurs binaires et fonctions coût

Nous avons testé 3 types de modèles : modèle linéaire **Logistic Regression** ; modèle ensembliste **Random Forest** « simple » ; et **LightGBM** « avancé », au sens où la construction des arbres de décisions est guidée par une méthode de gradient. Nous n'avons pas construit de fonction coût au-delà des fonctions de perte optimisant nativement ces algorithmes, respectivement ***log-loss*** en régression, minimisation de ***giny impurity*** et ***cross-entropy***).

Random Forest et **LightGBM** permettent également de valoriser des apprenants faibles et offrent des interprétations des « features importances ».

5. Méthodologie d'entraînement du modèle

L'ensemble des dossiers ou « applications » ayant été regroupés dans l'étape précédente, nous apprenons des « Application train » pour lesquels la cible est connue, puis nous effectuerons la prédiction finale sur les données « Application test » que nous étaierons via le tableau de bord.

Pour l'entraînement et la validation de nos modèles, nous procédons à une nouvelle séparation en données d'apprentissage et de test, selon une méthode de split stratifié selon la cible et une cross-validation sur 5 Folds pour nous assurer de la robustesse des résultats obtenus.

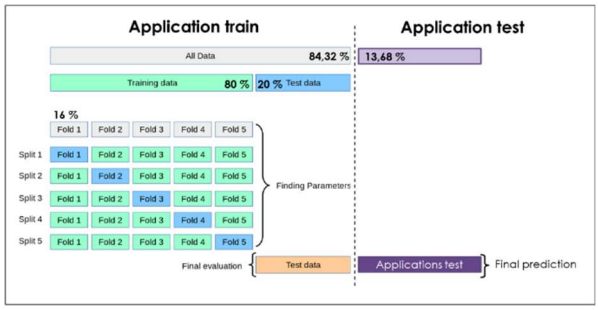


Figure 6 : Split et Cross Validation

Nous avons également testé une méthode de remédiation au problème de déséquilibre des classes, SMOTE, basée sur le rééquilibrage via la création de données de synthèse, que nous avons testé contre le paramètre des modèles qui permet de remédier au déséquilibre par une méthode de poids des classes.

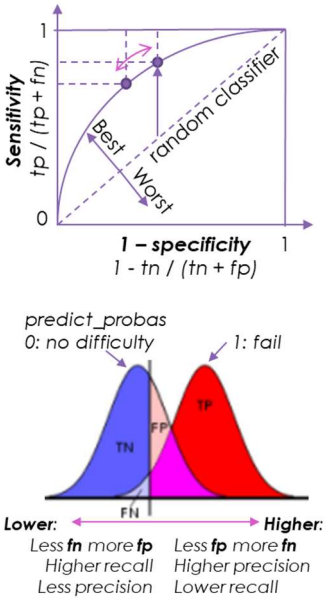
Cette méthode améliore la régression logistique, dégrade le Random Forest et n'a pas de plus-value pour le LightGBM.

6. Algorithme d'optimisation et métrique d'évaluation

Nous avons opté pour l'algorithme Hyperopt, afin d'explorer un large espace de recherche d'hyperparamètres des modèles, mettant en œuvre la cross-validation.

La fonction de coût retenue et appliquée en tant que ***« objective function »*** de notre algorithme Hyperopt est **le score AUC**, ou ***aire sur la courbe ROC.*** L'idée est d'obtenir un classificateur polyvalent, c'est-à-dire d'éviter d'optimiser nos paramètres pour un problème métier spécifique au détriment d'autres types d'optimisations que pourraient choisir le métier.

Nos modèles entrainés constituent une famille de classificateurs, représentés par leur **courbe ROC** illustrée ci-dessous (Receiver Operating Characteristic). Issue du traitement du signal, cette mesure est exploitée en classification binaire pour mesurer la performance du modèle : il est globalement d'autant plus performant qu'il maximise l'aire sous la courbe.



Figures 7a et 6b : Courbe ROC et notion de seuil de prédiction

La courbe ROC est obtenue en faisant varier le seuil de prédiction (ou discrimination, figure de droite), pour lequel la valeur de probabilité de classe 1 entraine une classe 1 prédite, et retourne les valeurs de sensibilité et d'antispécificité associées. Cela revient à moduler les effectifs de la matrice de confusion, illustrée ci-dessous.

Confusion Matrix	Actual 0 Good	Actual 1 Bad
Predict 0 Good	tn Usual business	fn High risk exposure
Predict 1 Bad	fp Loose clients	tp Usual business

Figure 8 : matrice de confusion

Ainsi, en fonction de la définition de notre cible :

- On s'attend à faire confiance à une prédiction négative en cas de **tn : true negative**, mais elle nous expose en cas de **fn : false negative** aux risques de défaut en accordant le prêt.
- On s'attend à faire confiance à une prédiction positive en cas de **tp : true positive**, mais elle signifie en cas de **fp : fasle positive**, refuser le prêt et perdre un « bon » client potentiel.

Nous avons besoin de minimiser les erreurs de prédictions avec un plus grand intérêt porté aux **faux négatifs**, par conséquence, minimiser les **fn** augmente les **fp.** Nous avons donc décidé d'observer une métrique alternative Fbeta où la valeur de beta représente l'importance relative accordée aux fp par rapport aux fn, pouvant traduire une métrique du dommage financier.

Un autre aspect très important dans notre contexte, est que notre classificateur étant non parfait nous aurons recours à une métrique supplémentaire afin d'optimiser les résultats des prédictions.

7. Meilleur modèle obtenu et simplification

Un **AUC score de 0.7646** est obtenu pour notre meilleur modèle : un **LGBMClassifier**, avec pour paramètres boosting_type='goss', class_weight='balanced', learning_rate=0.019…, max_depth=23, estimators=610, num_leaves=52.

L'utilisation du Fbeta score peut aider à identifier un meilleur seuil de prédiction, celui pour lequel le problème est optimal du point de vue métier.

$$F_\beta = (1+\beta^2)\cdot\frac{\text{precision}\cdot\text{recall}}{(\beta^2\cdot\text{precision})+\text{recall}}$$

With : . Precision = tp / (tp + fp)

. Recall = tp / (tp + fn)

$$F_\beta = \frac{(1+\beta^2)\cdot\text{true positive}}{(1+\beta^2)\cdot\text{true positive}+\beta^2\cdot\text{false negative}+\text{false positive}}$$

Figure 9 : Métrique additionnelle F-beta

Ci-dessous nous présentons les valeurs obtenues pour Fbeta (2) et le taux de défaut associés aux valeurs du seuil. Le seuil optimal au sens du Fbeta étant 0,47.

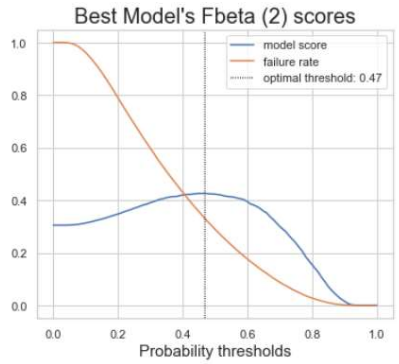


Figure 10 : Compromis sur le seuil de classification

Nous proposons de surveiller la valeur du taux de dossiers refusés, à positionner entre l'observation réaliste des taux de défauts constatés et le taux de refus des dossiers antérieurs. Ici le « meilleur » seuil représente plus de 30 % de dossiers rejetés ce qui paraît irréaliste même en considérant une part de dossiers annulés requalifiés en défauts évités.

➔ ***Une perspective serait de qualifier la métrique métier et confronter nos observations au contexte de la société.***

S'agissant de la simplification, nous avons retenus les features dont l'importance cumulée permet d'atteindre 80 % du total, cela permet de réduire de moitié le nombre de features requis (125) tout en conservant un score non dégradé : **l'AUC score obtenu est 0.764**.

➔ ***D'autres techniques de Features Selection peuvent être introduites pour tenir compte de l'intérêt combiné des features.***

➔ ***Une perspective pour le modèle serait de limiter des paramètres complexifiant son interprétabilité (de plus petits arbres de décisions).***

8. Interprétabilité du modèle

Malgré la qualité et la quantité d'informations sources disponibles et tous les efforts que nous avons déployés pour obtenir un modèle de Machine Learning efficace pour notre **Use Case** initial, nous obtenons un score médiocre en regard de l'objectif.

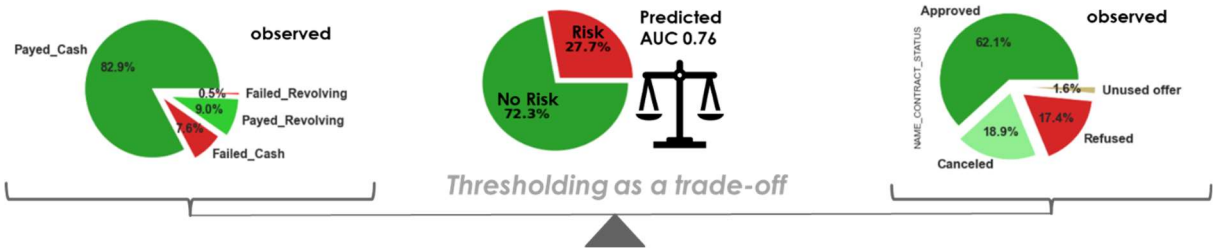


Figure 11 : compromis prédiction - observations, incluant la détermination du seuil

Par ailleurs, nous avons observé que l'optimisation d'un objectif métier revient à « *sévériser* » les décisions de refus de façon non réaliste. Faute d'élément complémentaire pour en juger, nous pouvons proposer les actions suivantes :

- Mener une analyse **locale**, pour exploiter le modèle pour sécuriser les accords de prêts, en priorité via l'analyse des dossiers présentant un faible écart de probabilité de défaut prédite, afin de proposer des ajustements (ex. sur les montants et durées engagées),
- Ajuster du seuil de prédiction de façon à s'approcher du taux de refus observé,
- Challenger les règles d'octroi en vigueur par rapport aux modalités de prédiction du modèle.

Ces propositions d'exploitation conduisent naturellement à approfondir la notion **d'interprétabilité** du modèle.

8.1. Model feature importance

Le premier pas dans l'interprétabilité d'un modèle consiste à observer l'importance relative des features.

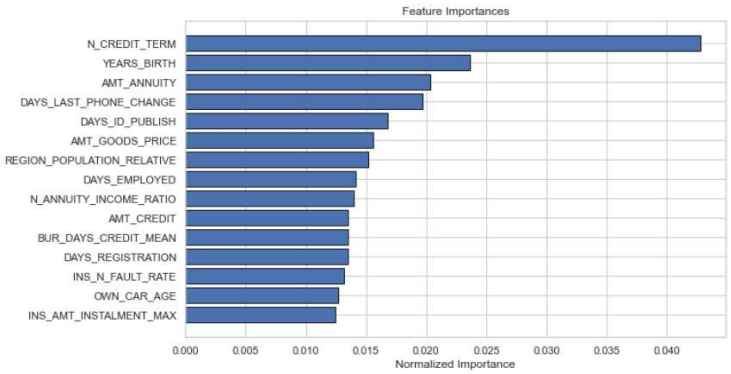


Figure 12 : LGBM most important features

Dans le cas de modèles d'arbres de décisions, l'importance d'un feature repose sur la ***Gini Importance*** ou ***Mean Decrease in Impurity (MDI)*** sommée à travers les arbres de décisions (nombre de nœuds) en proportion des effectifs séparés par ces nœuds. Ici, la durée du crédit (N_CREDIT_TERM) est le levier principal de notre modèle dans la prédiction de défauts et en conséquence dans la part de décision d'octroi d'un prêt. En variant ce terme (lié à l'annuité et au montant total), il est possible d'actualiser le résultat d'une prédiction Cette information ne permet toutefois pas d'analyser l'influence des features sur cette prédiction. Cela peut conduire à réduire la confiance des utilisateurs dans le modèle et ses prédictions.

➔ ***A ce titre une perspective serait d'agir sur les paramètres du modèle pour accroître la faculté d'interprétation d'un modèle : pour challenger des règles métiers.***

8.2. Technique SHapley Additive exPlanations[1]

Figure 13 : illustration technique SHAP

Dans le cas de SHAP, l'explainer prend le modèle en argument et propose des visualisations globales (summary) et locales. Cette technique est compatible des modèles d'arbre de décisions et s'appuie sur les valeurs de Shapley. La « force » d'un feature correspond à la moyenne, à travers toutes les permutations, des contributions d'un feature à la fonction de perte (coût).

Ci-dessous, une explication Force plot d'une prédiction, permet d'identifier la contribution locale des features. La prédiction est reportée manuelle et l'étendue est mesurée (f(x) = 0.79). On se place dans le cas d'antériorités pénalisantes avec un taux élevé de défaut de paiement sur les remboursements antérieurs.

Figure 14 Force plot pour une prédiction

SHAP permet également cette visualisation sur un ensemble, ici le sample est biaisé (plus de défauts prédits), mais la valeur n'est pas comparable à l'écart de probabilités prédites.

Figure 15 : Force plot pour un ensemble de prédictions

Ci-dessous pour le taux d'échéances avec défaut de paiement, l'importance du feature est d'autant plus grande que sa valeur est élevée, avec une force en proportion croissante pour la prédiction de la classe 1 : risque de défaut.

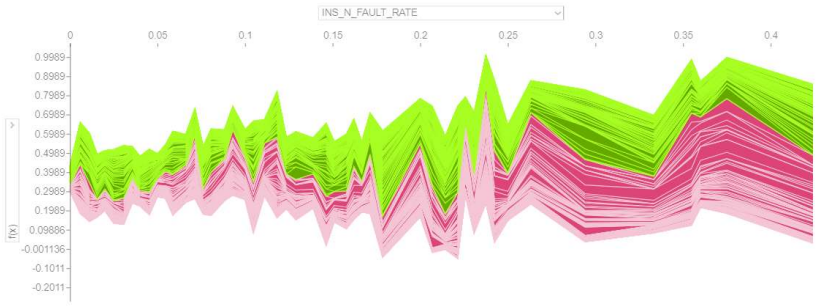


Figure 16 : Force plot – détail du taux de défauts de paiement - en considérant un sample des données

Ce type de visualisation est interactive. Il faut cependant changer de point de vue par rapport aux distributions en valeur habituelles, car on observe ici une distribution de « forces ».

[1] https://shap.readthedocs.io/

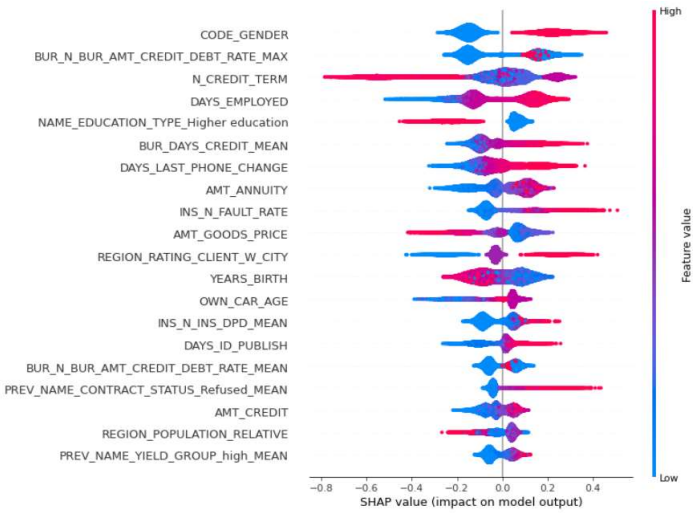


Figure 17 : Summary plot de la technique Shap

Enfin, une représentation caractérise l'ensemble du comportement du modèle (données sur lequel le fit est réalisé) et permet d'observer au niveau global quels sont les leviers les plus impactants.

Cela peut être considéré pour créer un ajustement global des modalités du prêt, mais cela ne permet pas d'identifier des leviers locaux propre à la demande en cours d'analyse.

8.3. Technique Local Interpretable Model-agnostic Explanations[2]

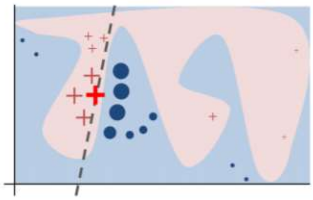


Figure 18 : illustration de la technique LIME

Dans le cas de LIME, l'explainer ne dépend que des valeurs (ici un classification binaire et l'utilisation de Lime.Tabular.Explainer). LIME fourni une mesure la contribution positive (***support***) ou negative (***contradict***) de chaque feature à l'appartenance à la classe prédite, mesurée via la distance et le poids des voisins. Un intérêt au-delà de notre contexte est de traiter les données textuelles et images.

L'illustration ci-dessous présente la vue intégrée au notebook et résumant la prédiction (les probabilités sur les 2 classes dans notre cas. Ainsi, à l'instar des valeurs Shap, de la contribution des features à la prédiction (les 10 premiers en importance relative).

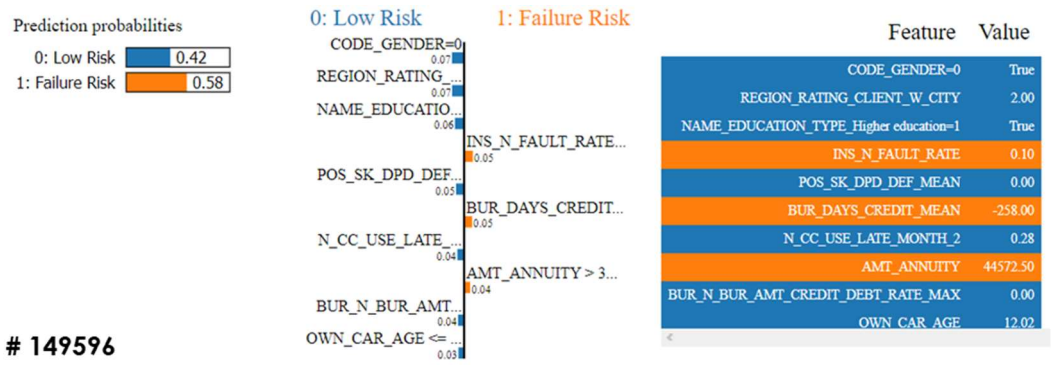


Figure 19 : explication LIME

[2] Utilisation https://lime-ml.readthedocs.io/; base théorique https://arxiv.org/pdf/1602.04938.pdf

Nous avons privilégié la production de bar plot simples reprenant les informations utiles, comme le montre la figure ci-dessous. Comme il s'agit de la prédiction de Classe 1 en risque, nous avons choisi de conserver la couleur rouge pour la contribution positive : ***support*** de chaque feature à la prédiction, cela revient au sens de lecture de la technique SHAP. Nous avons repris et modifié la partie de code open source du package LIME sur Github pour modifier ce graphe et faciliter les explications.

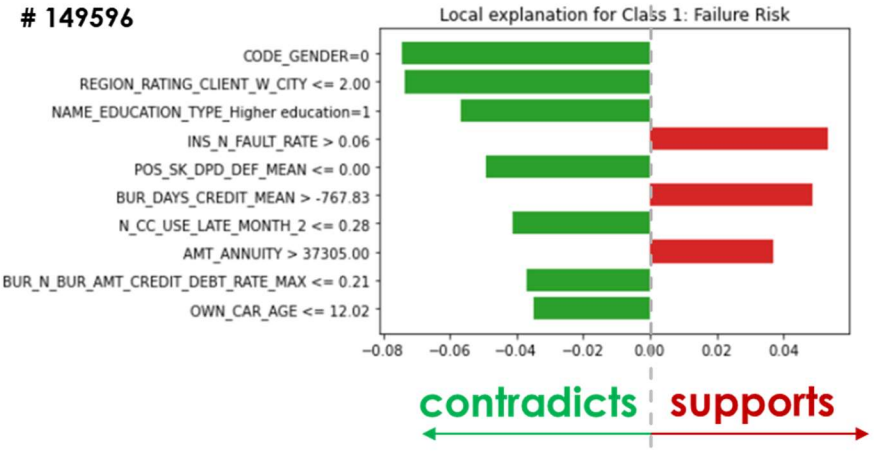


Figure 20 : notre choix de représentation en technique LIME

L'intérêt de cette visualisation est qu'elle nous a permis de décliner facilement un comparatif des valeurs de l'application par rapport à la moyenne des dossiers similaires (que nous avons réalisé sur un voisinage au sens des features les plus importants), ainsi que par rapport à la moyenne des 2 classes 0 ou 1.

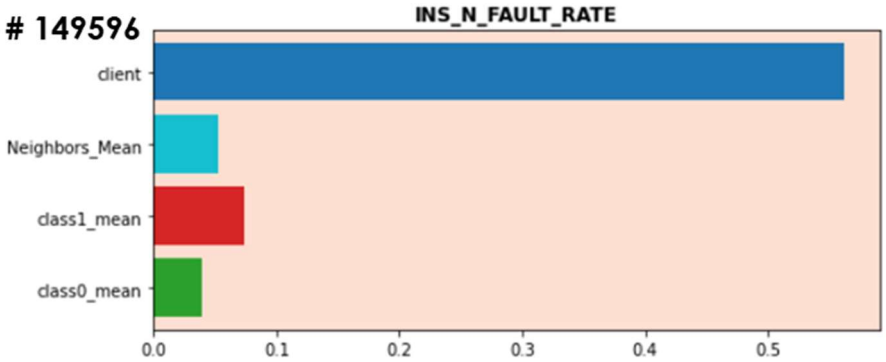


Figure 21 : Focus sur une 'anomalie' expliquant la prédiction Class 1 : Risque

En complément de la moyenne des prédictions pour les dossiers similaires, cela permet de détecter des valeurs en écart. Ici le niveau du taux de défauts sur l'antériorité des versements est rédhibitoire, malgré des valeurs favorables pour les 3 premiers features en ordre d'importance.

- ➔ ***Au vu des concepts sous-jacents et confirmé par nos observations, avec par ailleurs des paramètres propres à chaque technique, il ne nous semble pas possible de mener de façon complémentaire les deux investigations.***
- ➔ ***Il convient de préciser l'objectif d'interprétation et de confirmer la technique appropriée.***

- ➔ ***Le plus grand champ d'action concerne la façon de considérer les features eux-mêmes dans la mesure où il faut séparer ceux qui expliquent (ex. le genre H/F), causent (l'antériorité de défauts) ou modulent (le ratio annuité / salaire) une prédiction***