DATA

Square

DONNER DU FUTUR AU TALENT

OPEN DATA : ACCÉLÉRATEUR DE BUSINESS

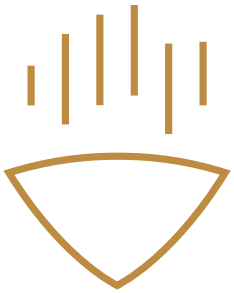
SAMY FRACSO, ADRIEN AUBERT, MICHAËL TOUZE, ALEXANDRE ZAVAGNO, NICOLAS DERRIEN, HUGO BATACHE

FOCUS I MARS 2022

Square

DONNER DU FUTUR AU TALENT

DATA



OPEN DATA : ACCÉLÉRATEUR DE BUSINESS

SAMY FRACSO, ADRIEN AUBERT, MICHAËL TOUZE, ALEXANDRE ZAVAGNO, NICOLAS DERRIEN, HUGO BATACHE

SOMMAIRE

Square

DONNER DU FUTUR AU TALENT

PRÉAMBULE

S'il est convenu que la donnée représente une nouvelle forme de patrimoine pour l'entreprise et constitue à ce titre un moyen d'activation de leviers de croissance, il est également admis que son potentiel est encore loin d'être totalement appréhendé, et ce en raison de plusieurs difficultés aussi bien empiriques qu'historiques : infrastructures techniques pas toujours mises à niveau, lacunes culturelles et rareté des compétences, etc.

L'Open Data constitue un moyen d'améliorer cette situation. L'Open Data désigne l'ensemble des données numériques libres de droit. Elle repose sur deux principes clés : le partage (les données doivent être accessibles gratuitement par tous) et la réutilisation (les données doivent être facilement exploitables et réutilisables).

Dans le secteur bancaire, les Open Data restent très peu utilisées alors qu'elles représentent une source indéniable d'informations qui peuvent apporter des analyses clés en complément des données internes. Il est vrai que plusieurs incertitudes et contraintes entourent l'utilisation des Open Data : gains de performances incertains, charge significative d'adaptation des processus et de montée en compétences, transparence et qualité des données, etc.

Les freins à contourner peuvent paraître nombreux ; l'objet de ce nouveau Focus Square est d'illustrer les gains possibles à recourir à l'Open Data dans les projets opérationnels que peuvent mener les établissements bancaires. Après une brève présentation de l'Open Data, nous nous concentrerons sur les avantages et difficultés couramment rencontrés dans leur utilisation puis dans un troisième temps nous tenterons de dépasser ces freins en prouvant l'efficacité des Open Data à travers 2 cas d'usages.

1.
PRÉSENTATION DE L'OPEN DATA

Avant de rentrer dans le détail des avantages et limites de l'Open Data dans le cadre d'usages concrets, nous proposons dans ce chapitre de rappeler la définition et le contexte du développement de l'Open Data ; nous évoquerons également les différentes initiatives publiques de soutien à l'Open Data.

CADRE GÉNÉRAL

L'Open Data couvre l'ensemble des données numériques libres de droits. Elle repose sur deux principes clés : le partage (les données doivent être accessibles gratuitement par tous) et la réutilisation (les données doivent être facilement exploitables et réutilisables). Si, au niveau international, de nombreuses typologies de licences ouvertes cohabitent, deux licences sont aujourd'hui applicables en France pour l'information publique :

- **La Licence Ouverte (LO)** : spécificité française et proposée par Etalab, il s'agit de la licence la plus libérale, octroyant à l'utilisateur les droits suivants sur la source de données :

• de la reproduire, la copier,

• de l'adapter, la modifier, l'extraire et la transformer, pour créer des « Informations dérivées », des produits ou des services,

• de la communiquer, la diffuser, la redistribuer, la publier et la transmettre,

• de l'exploiter à titre commercial.

Avec pour seule contrepartie de mentionner l'auteur originel de la base de données.

- **L'Open DataBase Licence (ODbl)** : licence internationale dérivée du droit anglo-saxon, elle octroie également des droits de reproduction, copie, adaptation, modification ; la différence majeure avec la Licence Ouverte repose sur l'obligation de restitution de la donnée sous la même licence ODbl (sauf contrepartie financière).

Pour en savoir plus : Licences - data.gouv.fr

L'AVANCE FRANÇAISE

La mise à disposition de ces données est assurée à la fois par des organismes privés mais aussi publics. En France, le gouvernement a créé en 2011 Etalab, une administration[1] qui « coordonne la conception et la mise en œuvre de la stratégie de l'État dans le domaine de la donnée » (https://www.etalab.gouv.fr/qui-sommes-nous). Etalab a lancé la même année le site data.gouv.fr qui répertorie l'ensemble des données publiques mises à disposition par les différentes administrations françaises. On y trouve des données sur un grand nombre de secteurs : culture, santé, économie etc.

En 2016, la loi 2016-1321 du 7 octobre pour une République Numérique est venue confirmer l'effort ainsi que l'avance française dans le domaine de l'ouverture des données publiques. Cette loi a modifié effectivement les modalités de l'accessibilité de la donnée publique, l'Open Data devenant la norme et non plus l'exception. L'ensemble des services publics sont concernés : administration d'état, collectivités locales de plus de 3 500 habitants, établissements publics ainsi que les organismes privés chargés d'un service public.

Entre autres, l'article 14 de la loi pour une République Numérique promeut la création d'un ensemble de données dites de référence, identifiées comme les données ayant le plus fort impact économique et social. Parmi les 9 sources identifiées à ce jour, deux nous serviront notamment de trame au développement et à la mise en avant du potentiel économique des Open Data :

- La base Sirene des entreprises et de leurs établissements

- La base Code Officiel Géographique (COG)

La liste des données de référence est consultable sur le lien suivant :
https://www.data.gouv.fr/fr/reference

Autres avancées apportées par la loi Numérique, l'arrivée de nouvelles données dites d'utilité générale :

- Délégation de service public (énergies, déchets) ;

- Domaine juridique ;

- Mobilité & transport (confirmé par la loi de 2019 sur l'orientation des mobilités, en cours d'application).

Ces initiatives du gouvernement placent la France au 4ème rang de l'Open Data Index (1er européen), un indice créé en 2013 par l'Open Knowledge Foundation permettant de mesurer et de classer les pays en matière d'ouverture et d'accessibilité de la donnée. Cette association anglaise née en 2004 a vocation à promouvoir la circulation, l'accessibilité à l'information, pour tous et gratuitement. Classement complet : https://index.okfn.org/place/

Deux axes pénalisent la performance française (derrière Taiwan, l'Australie et la Grande Bretagne) : l'ouverture des données sur les dépenses de l'Etat (sur une maille acheteur / vendeur), ainsi que l'accessibilité à un cadastre complet, aux impôts et prix de vente. Sur ce dernier point, la mise à disposition publique de la base Demande de Valeur Foncière (DVF) depuis 2019 doit permettre une réévaluation positive de la position de la France.

1: Etalab est une direction de la DINUM (Direction Interministérielle du Numérique)

L'OPEN DATA JURISPRUDETIELLE

L'ouverture des données dans le domaine juridique, initialement prévue dans la loi pour une République Numérique, a été confirmée par la loi du 23 mars 2019 de programmation et de réforme de la justice.

Cette loi a notamment eu pour impact de rendre accessible et gratuit l'ensemble des jugements, sous réserve du respect de la vie privée des parties prenantes. Révolution dans un secteur où la reprise des historiques de jugements et des condamnations constituait une tâche longue et incompressible, retrouvez la publication Square à ce sujet :

Justice et intelligence artificielle : les algorithmes sont-ils nos juges de demain ? (journaldunet.com)

EXEMPLE D'UTILISATION

Dernier succès majeur en date pour l'utilisation de l'Open Data : CovidTracker.fr. À la suite de la crise sanitaire du Covid-19, ce site internet intègre des statistiques et graphiques sur l'évolution de la pandémie en France et dans le monde. En 2021, une section sur le suivi de la vaccination est intégrée, on y trouve par exemple la répartition des vaccinés par région ou encore la localisation des centres de vaccination. Facile d'utilisation, avec des graphiques clairs et dynamiques, ce site est devenu une référence en France sur le suivi de la pandémie et est également utilisé par le gouvernement.

Intégralement construites à partir d'Open Data, les données utilisées proviennent, sur le plan national, de Santé Publique France et de l'INSEE (Institut National de la Statistique et des Etudes Economiques). Les analyses sur le plan international sont issues des données mises à disposition par le CSEE (The Center for Systems Science and Engineering) de la John Hopkins University. Dans une logique d'open source, l'auteur a également rendu public l'ensemble de ces programmes essentiellement développés en langage Python.

2.

L'OPEN DATA, UN FACTEUR DE PERFORMANCE POUR LES INSTITUTIONS FINANCIÈRES

CONTEXTE

Les données sont omniprésentes dans l'activité des entreprises et sont désormais utilisées tout au long du cycle de vie des opérations courantes menées avec les clients et les fournisseurs. Ces données permettent aux institutions financières de développer des systèmes de recommandation, des modèles de notation de la performance, d'évaluation des risques financiers, de tarification, d'appétence commerciale ou de fraude. C'est dans l'exploitation de ces données, majoritairement voire exclusivement d'origine interne, que réside une grande partie de la performance.

Dans ce cadre, les Open Data ouvrent de nouvelles possibilités et des applications innovantes permettant de parfaire et varier les services proposés aux clients.

DE NOMBREUSES OPPORTUNITÉS DANS LE SECTEUR FINANCIER

Dans une économie de plus en plus data-centrée, les institutions financières peuvent être considérées comme les pionnières de l'utilisation des sciences des données avec l'intégration d'outils de scoring dans les années 1950.

L'initiative française de la loi numérique a mis à disposition les bases de données Sirene, INSEE et Banque de France. Ces dernières représentent une somme d'informations qui permettent de dériver de nouvelles analyses et une meilleure vision de la santé financière de l'entreprise. Entre autres, ces données rendent possible l'étude de la cessation de paiement (cf. notre premier cas d'usage) et l'analyse de l'activité sectorielle de manière approfondie sous de nombreux prismes : par le biais des données géographiques, il devient possible d'observer l'essor ou le déclin de certains secteurs d'activité à différents niveaux (région, ville, etc.).

De nombreuses informations dérivent également de ces données publiques : les bases Sirene contiennent l'ensemble des établissements associés à une entreprise ce qui permet d'observer les dates d'ouverture, de fermeture et les adresses de ces derniers.

La consolidation et le croisement de ces informations peuvent alors constituer des faisceaux d'indicateurs convergents dans la détection de comportements atypiques, à l'origine de la lutte contre les différentes formes de fraudes.

De plus, si le phénomène de l'Open Data est initialement apparu suite à l'étude du climat et à la publication des informations utilisées dans ces analyses, il existe dorénavant des données structurées permettant d'introduire de nouvelles notions dans les modèles de notation. Pour ne citer qu'un exemple, les acteurs de l'immobilier de Chicago partagent leurs données de consommation d'énergie, il devient alors possible dans ces zones d'identifier les entreprises les plus sujettes à la taxe carbone.

UNE AMÉLIORATION DE LA RÉPONSE AUX EXIGENCES RÉGLEMENTAIRES

Dans un contexte réglementaire de plus en plus contraignant, la BCE a publié en mai 2020 un guide sur la gestion des risques climatiques qui spécifie les attentes de la supervision concernant les modalités de prise en compte de ces risques autant dans la gouvernance que dans la gestion des risques. Il est ressorti des premières études (notamment via l'exercice pilote de l'ACPR publié en 2021) que l'enjeu principal de cette analyse réside dans la récolte des données nécessaires et il y a fort à parier que ce manque sera en partie comblé par les Open Data puisque certaines données sont déjà disponibles et structurées.

D'un autre point de vue, les données externes peuvent aussi améliorer les études de la qualité de la donnée grâce à la mise en place de contrôles de fiabilisation. Dans le cas de la base Sirene, il devient possible de comparer le code d'activité APE récolté par l'institution à celui disponible dans ces bases de données publiques.

DE NOUVELLES SOURCES DE DONNÉES POUR DES MODÈLES PLUS PERFORMANTS

Bien que les modèles statistiques d'estimation du risque de crédit puissent être considérés comme proprement dimensionnés, l'introduction de typologies d'informations non utilisées sont presque systématiquement signe de gain de performance. En conservant l'exemple des données climatiques ; la littérature[2], à maintes reprises, a montré le lien entre les risques financiers et l'empreinte carbone. En ce sens, les données climatiques libres de droit représentent donc une ressource non exploitée pour l'amélioration de la performance des modèles existants.

UNE MEILLEURE CONNAISSANCE CLIENT – DES SERVICES PERSONNALISÉS

L'Open Data permet aussi d'améliorer la connaissance client : puisqu'une institution financière dispose des informations de paiement de ses clients, elle peut dorénavant catégoriser ces derniers par le biais des bases Sirene en rattachant l'entreprise vers laquelle le paiement a eu lieu à son secteur d'activité (cf. notre deuxième cas d'usage). Cette étude ouvre de nombreuses possibilités comme l'identification de signes précoces de dégradation de la santé financière par une vision pertinente des ratios revenus/dépenses ou la détection d'un profil client plus fin avec l'étude de ses opérations.

2: https://www.sciencedirect.com/science/article/abs/pii/S0959652620316814?via%3Dihub

La catégorisation des paiements permet également de proposer une synthèse pertinente des opérations et même d'extrapoler une optimisation des dépenses et recettes : il est par exemple tout à fait possible de planifier le paiement des factures en fonction des flux financiers, ou même en effectuant des virements automatisés entre les comptes épargnes et courants. De nombreuses autres améliorations du service client pour une meilleure personnalisation de la relation peuvent ainsi être envisagées.

DÉVELOPPER DE NOUVEAUX SERVICES POUR LES CLIENTS PROFESSIONNELS ET ENTREPRISES

Les exemples vus précédemment permettent aussi d'anticiper de nouveaux services pour les professionnels : l'analyse sectorielle permet d'identifier les territoires susceptibles de générer croissance et emploi. En effet, il est possible d'identifier les zones où certains services sont en pénurie ainsi que de mettre en relation différents acteurs pour optimiser les chaines de productions, vente, transport, etc.

Il est aussi possible de combiner cette analyse sectorielle avec un modèle affiné de cessation d'activité (par secteur) pour améliorer les études portant sur la conquête de territoire. En combinant ces informations aux outils de visualisation de données, il devient possible de construire un outil supplémentaire d'aide à la décision en livrant aux experts métiers des tableaux de bord interactifs.

En somme, les données étant d'ores et déjà au cœur du cycle de vie d'un produit financier, l'enrichissement de ces dernières permet une amélioration des processus existants en apportant de nouvelles possibilités d'analyse. Néanmoins, force est de constater que le secteur financier peine à mettre en production des solutions utilisant ces nouvelles ressources. Cet effet s'explique en partie par les challenges inhérents à l'exploitation de ces informations.

3.

L'OPEN DATA, DES FREINS À SURMONTER POUR LES INSTITUTIONS FINANCIÈRES

Malgré le développement des Open Data ces dernières années, plusieurs obstacles empêchent les banques et les assurances d'adopter pleinement leur utilisation. Certaines caractéristiques intrinsèques aux Open Data peuvent en effet représenter un frein à leur exploitation.

PRINCIPAUX FREINS INHÉRENTS À LA NATURE MÊME DES OPEN DATA

Les données ouvertes étant diffusées par des organismes tiers, la qualité et la confiance dans l'information fournie sont des éléments essentiels de développement de leur utilisation. Des problèmes de véracité, les incertitudes sur la source à l'origine de l'information et un manque de cohérence de la donnée dans le temps peuvent freiner les acteurs privés à adopter leur utilisation, et les pousser à privilégier uniquement l'utilisation des informations internes sur lesquelles ils peuvent avoir une plus grande confiance.

En outre, même si la qualité des données peut être assurée par des analyses et des contrôles supplémentaires, la compréhension des données fournies peut s'avérer complexe. C'est dans la variété et la diversité que réside la valeur des Open Data, mais la description et les caractéristiques de chaque variable (méta-data) fournies par la source peuvent s'avérer insuffisantes ou ambiguës, devenant ainsi source d'erreurs d'interprétation et donc pénalisant in fine les résultats espérés. Le recours aux Open Data nécessite une connaissance pointue de la donnée exploitée.

Les Open Data peuvent également souffrir d'un manque d'uniformité et de standardisation. En effet les formats, les fréquences de rafraîchissement et les structures utilisées varient en fonction de la source et cela peut constituer un obstacle à la combinaison des différentes informations. L'interopérabilité entre les sources Open Data et les données internes à l'entreprise ou entre les sources Open Data entre elles, complique la conduite opérationnelle des projets. Ce manque de standardisation peut engendrer un effort et des coûts techniques supplémentaires afin d'adapter des systèmes d'informations existants aux différentes sources Open Data. La mise en production de ces données peut donc également devenir un véritable frein.

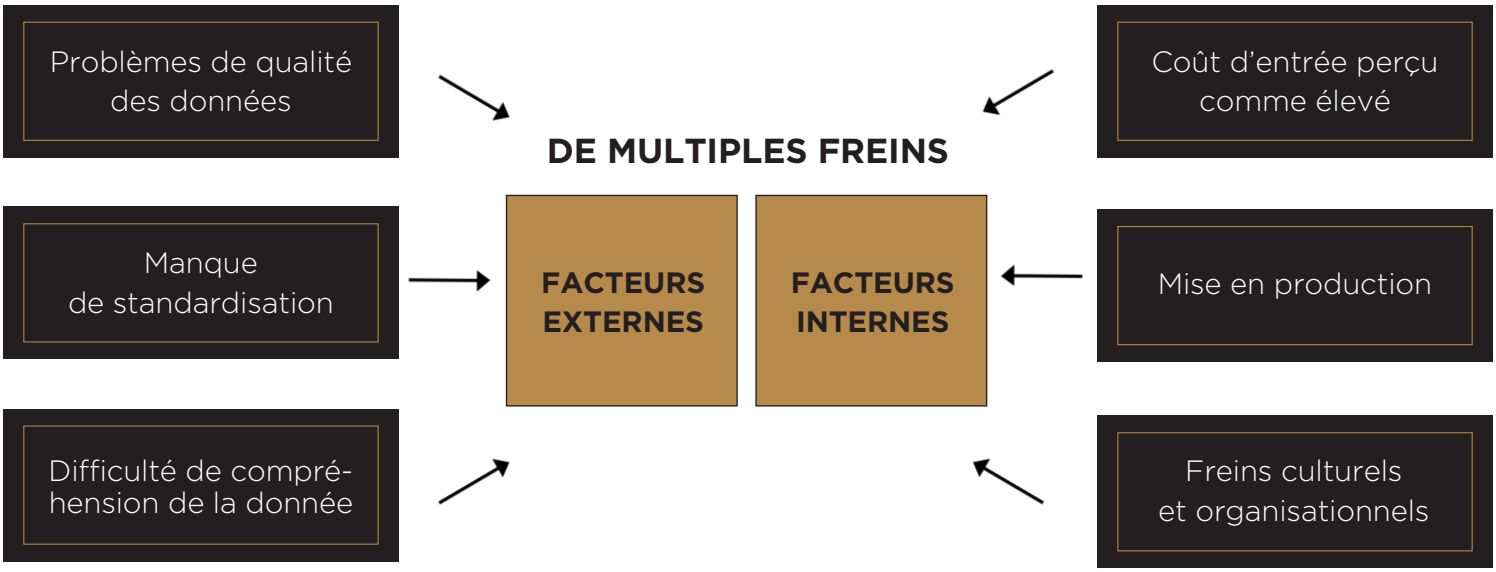
UNE CULTURE À DÉVELOPPER

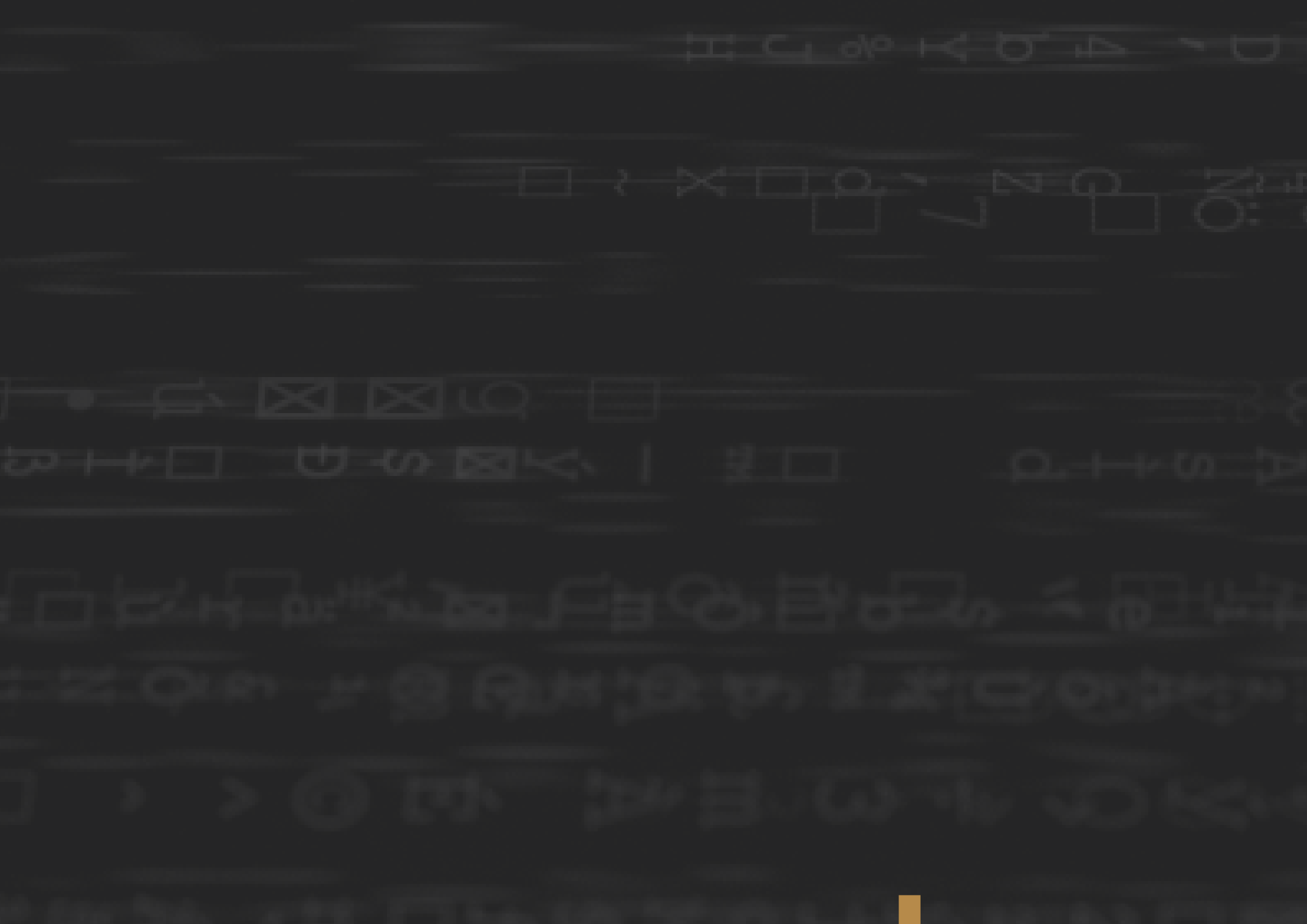
Comme évoqué, intégrer les Open Data de façon pérenne en dehors du simple prototype (proof of concept) nécessite l'intervention d'une multitude de compétences, autour de la collecte automatisée, l'harmonisation et l'homogénéisation au sein d'un modèle de données partagé, le contrôle de la qualité et l'analyse de l'information. Tous les métiers gravitant autour de la data sont concernés, de l'IT aux Data Scientists, en passant par les data managers et analystes métiers, ce qui peut ainsi être perçu comme un investissement coûteux et chronophage.

Or, bien souvent ces compétences sont déjà présentes dans les entreprises et les problématiques spécifiques aux Open Data ne sont pas si éloignées de celles qui existent déjà pour les données internes. En effet, les institutions financières se sont dotées ces dernières années d'un pôle Data en charge de la gouvernance de la donnée, du reporting, du pilotage de la qualité de la donnée et du développement des infrastructures au service de l'ensemble des parties prenantes de l'entreprise.

Les institutions financières s'appuient principalement sur les données collectées en interne et sur leurs systèmes d'information matures. Et du fait de la disponibilité des données internes, le recours aux Open Data n'est pas systématiquement évoqué dans les premières phases des projets data. L'organisation et les processus existants dans ces structures n'incitent pas à la recherche et l'exploration de sources de données externes et libres. De plus, la segmentation des compétences dans les projets data où chaque intervenant possède une expertise propre et n'intervient pas dans toutes les phases du projet (équipes en charge de la collecte, équipes chargées de l'analyse, équipe métiers) induit à utiliser l'existant et n'encourage pas à aller chercher de nouvelles sources de données.

L'utilisation des Open Data demande par conséquent un effort de prospection et de veille à placer au cœur de la stratégie et de la culture data. Finalement les contraintes intrinsèques peuvent constituer un frein à leur exploitation mais, plus que tout, ce sont les habitudes culturelles et organisationnelles des entreprises qui peuvent être le premier obstacle à l'exploitation des Open Data.


Problèmes de qualité des données
Manque de standardisation
Difficulté de compré-hension de la donnée
DE MULTIPLES FREINS
FACTEURS EXTERNES
FACTEURS INTERNES
Coût d'entrée perçu comme élevé
Mise en production
Freins culturels et organisationnels



4.

USE CASE : DÉTECTION DE LA CESSATION D'ACTIVITÉ DES ENTREPRISES

CONTEXTE & ENJEUX

La détection des signes précoces de la défaillance d'entreprise reste un enjeu majeur pour de nombreux acteurs économiques tels que les organismes de financement, les investisseurs privés et institutionnels ainsi que les organismes publics. Pour les banques, il s'agit dorénavant d'une exigence réglementaire à travers l'exigence "Loan Origination & Monitoring" issue de l'Autorité Bancaire Européenne (EBA en anglais), qui promeut ici la mise en place des Early Warning Indicators (EWI), c'est-à-dire le traitement de signaux faibles susceptibles de capter la dégradation de la solvabilité d'une contrepartie.

Les PME constituent le principal maillage du tissu économique du territoire, l'évaluation du risque de défaillance des entreprises via des modèles de notation est une pratique particulièrement répandue principalement dans les établissements de crédit et les organismes de notation. D'ordinaire, le principe consiste à extrapoler les informations descriptives, comportementales et financières des différentes entreprises pour établir un modèle de probabilité de défaillance global.

Depuis la mise à disposition des Open Data WebStat (Banque de France), Sirene (INSEE) et des Greffes des Tribunaux de Commerce (Data Infogreffe), un champ d'exploration s'est ouvert dans ce domaine avec la possibilité d'exploiter à la fois les données caractéristiques des entreprises et un ensemble hétérogène d'informations comportementales et financières agrégées par secteur d'activité, typologie d'entreprise ou zone géographique.

OBJECTIF

L'objectif de ce cas d'usage consiste à exploiter les données financières et comportementales agrégées par secteur d'activité principalement pour établir un modèle de classement capable de détecter avec fiabilité, le risque de cessation d'activité. Ce modèle peut permettre d'améliorer les performances des modèles de risque déjà existant, notamment pour les structures financières avec un nombre de clients Professionnels et Entreprises trop limité pour réaliser un modèle pertinent avec uniquement des données internes.

MÉTHODOLOGIE & RÉSULTAT

La notion de cessation est prise dans le sens large d'entreprises qui cessent leur activité et cessent d'exister juridiquement. La cessation d'une entreprise correspond à la fin de vie d'une entité juridique. Pour tenir compte à la fois des impératifs juridiques et des réalités économiques, deux catégories de cessation sont prises en compte dans ce use-case :

- La cessation juridique de l'entreprise : une entreprise est cessée en cas de dissolution s'il s'agit d'une personne morale, et en cas de décès ou lors de la cessation de toute activité s'il s'agit d'un entrepreneur individuel ;

- La cessation d'activité de l'entreprise : l'entreprise cesse son activité. On parle aussi de cessation économique lorsque tous les établissements de l'entreprise sont fermés.

Ainsi, cette situation survient à l'issue d'une procédure de cessation d'activité d'une entreprise et aboutit à la fermeture et à l'arrêt de son activité. Il peut s'agir d'un dépôt de bilan ou d'un arrêt volontaire. Dans tous les cas, l'entreprise est passée subséquemment par les étapes de dissolution, liquidation, paiement des impôts et déclaration sociale.

La cessation ne doit pas être confondue avec la défaillance d'entreprise. On parle de défaillance lorsqu'une entreprise fait l'objet de l'ouverture d'une procédure de redressement judiciaire à son encontre. Toutes défaillances n'induisent pas de cessations.

Différents algorithmes ont été testés :

- Un modèle expert basé sur une régression logistique où l'ensemble des choix de modélisations et métriques observées respectent une cohérence métier et une logique financière (par exemple en s'assurant que le taux de cessation augmente quand la dette accumulée augmente).

- Plusieurs modèles de machine learning (Decision Tree, Random Forest, GBM, etc.), affranchis de tout a priori métier sur la phase de features engineering, à des fins de benchmark et d'exploration.

FEATURES ENGINEERING

Le features engineering représente l'étape de sélection et de transformation de la donnée brute en information afin d'en maximiser sa pertinence lors du processus de modélisation.

Cette phase est cruciale dans un projet data science :

- Elle apporte de nouvelles dimensions d'analyse et d'explicabilité,

- Elle est déterminante dans la performance et la stabilité du modèle, au même titre que la phase de raffinement des modèles.

Elle est cependant parfois sous-estimée par les data scientists au détriment de la technique pure. C'est l'étape la plus chronophage d'un projet, et celle nécessitant le plus de connaissances métier.

DONNÉES UTILISÉES

INSEE / SIREN : données renseignées au niveau « Unités Légales » sur l'ensemble de l'historique (date de création, état administratif dont cessation, activité principale, etc.).

Banque de France (Webstat) : informations liées au bilan et compte de résultat disponibles à différents niveaux d'agrégation tels que le secteur d'activité, la zone géographique et la périodicité. Par exemple : la capacité d'endettement, le poids des capitaux propres appelés dans les ressources, la part des dettes bancaires dans l'endettement financier, etc.

MODÉLISATION – MODÈLE EXPERT

Régression logistique : après avoir catégorisé et appliqué des méthodes de sélection de variables, le modèle expert est construit autour des inputs suivants :

- Part des dettes bancaires dans l'endettement financier

- Productivité de la main d'œuvre (rendement apparent : valeur ajoutée en volume sur nombre d'heures travaillées)

- Poids des capitaux propres appelés dans les ressources (Variation entre deux arrêtés N/N-1)

- Taux de marge (Variation entre deux arrêtés N/N-1)

- Variation du taux de défaillance régionale générationnelle (Année N/N-1)

- Ancienneté depuis le dernier changement d'activité

- Nombre d'établissements fermés associés à l'unité légale.

Résultat : le modèle présente d'excellentes performances sur les métriques de précision (95%) et de rappel (96%) pour les établissements actifs. Pour les unités légales en cessation d'activité, le rappel (76%) est également très correct compte tenu de la structure des données et du faible nombre de variables caractéristiques disponibles.

Ce modèle prouve donc sa capacité d'apprentissage sur la détection des cessations d'activité à partir d'informations financières agrégées par secteur d'activité.

Les caractéristiques retenues sont hautement significatives en particulier d'un point de vue financier. En effet, le risque de cessation d'activité d'une unité légale dans l'année qui suit la dernière publication des résultats est d'autant plus élevée lorsque :

- La part de la dette bancaire dans l'endettement financier est supérieure à 35% ;

- La productivité apparente du travail (rendement apparent : valeur ajoutée en volume sur nombre d'heures travaillées) est inférieure à 66% ;

- La variation annuelle des capitaux propres est inférieure à -2% ;

- La variation annuelle du taux de marge est inférieure à 1%.

La maîtrise de l'endettement bancaire garantit la maîtrise de l'endettement financier net au bilan et permet de se prémunir de l'accroissement des exigibilités à court terme ainsi que de l'accroissement du poids des intérêts financiers dans le résultat. Par ailleurs, la maîtrise du taux de marge, la maîtrise des capacités de production ainsi que la variation des capitaux propres ont un impact significatif sur la pérennité de l'activité des PME.

MODÉLISATION – MODÈLES CONCURRENTS

Dans le but d'identifier les méthodes de classification qui permettraient d'optimiser la capacité de détection du risque de cessation d'activité, les résultats de 9 méthodes de classification ont été évalués (K-Nearest Neighbors, Linear SVM, RBF, SVM, Decision tree, Random Forest, Neural Net, Gradient Boosting, Naive Bayes, Quadratic classifier).

Les méthodes K-Nearest Neighbors, Decision Tree, Random Forest et Gradient Boosting Tree produisent les résultats les plus satisfaisants avec un léger avantage pour la Random Forest et le Gradient Boosting Tree avec un rappel (unités légales en cessation d'activité) respectivement à 92% et 93% vs 76% pour le modèle de régression logistique.

Il existe donc une différence significative entre les résultats obtenus à l'issue du modèle expert (régression logistique) et ceux obtenus à l'aide des modèles concurrents. Cette différence réside principalement dans la sélection des features. Si la part de la dette bancaire dans la dette financière, la variation du taux de marge et la productivité apparente du travail sont des facteurs communs aux deux méthodes, la plupart des variables descriptives à l'exception du nombre d'établissements fermés sont exclues par les modèles benchmarks.

L'ESSENTIEL À RETENIR

Les facteurs de risque discriminant définis à l'issue de l'étape de feature sélection supervisée de la cessation d'activité ont permis d'obtenir des segmentations des facteurs pertinents comparables à ceux communément utilisés dans le cadre de l'analyse financière.

Concernant les modèles challengers, les analyses démontrent qu'il existe une différence significative dans l'affectation des risques de cessation selon la méthodologie de classification et le secteur d'activité en particulier. En conséquence, l'analyse des données issues de l'Open Data sont de nature à permettre d'identifier des axes pertinents aux fins de l'analyse du risque sectoriel externe. Cette méthode peut notamment être utilisée dans le cadre de la détermination du risque à l'acceptation, dans le but de caractériser le risque relatif au dossier sur lesquels aucune observation du risque n'est disponible en interne.

Bien que les données Sirene ne comportent que peu de facteurs de risque comportementaux, l'évaluation du changement de secteur d'activité, de caractère employeur de l'unité légale ou encore du changement de statut juridique traduisent des changements structurels des entreprises, et sont par conséquent des signes d'amélioration ou de détérioration probable de la qualité des entreprises.

5.

USE CASE : LABELLISATION DES OPÉRATIONS BANCAIRES

CONTEXTE ET ENJEUX

Nous avons tous été déjà confrontés, lors de la lecture de notre relevé d'opérations bancaires, à une incertitude sur un achat carte : un libellé peu explicite, un nom de commerçant inconnu, une zone géographique qui ne correspond pas. Il s'agit le plus souvent de noms tronqués voire de sigles qui ne permettent pas de reconnaître immédiatement le nom de l'enseigne en question. Charge alors au client d'utiliser des outils adéquats pour permettre l'identification du bénéficiaire du paiement en cas de doute sur une transaction, ce qui se traduit le plus souvent par une perte de temps.

Du côté des établissements bancaires, ces incertitudes et cette absence de clarté ne constituent en aucun cas un avantage vis-à-vis de leurs clients : le manque de structuration de l'information contenue dans les opérations ne permet pas d'analyse fine et par conséquent d'améliorer de façon notable la connaissance client.

Qu'en est-il si cette information devenait structurée ? Avec un identifiant unique du commerçant, une localisation, un secteur d'activité associé à chaque opération permettrait de dégager une multitude de cas d'usage :

- Marketing Client : clustering plus fin des clients en lien avec leur poste de dépense, leur secteur préférentiel, la localisation de leurs dépenses. Plus largement cela pourrait permettre une segmentation dynamique, grâce à l'identification de nouveaux comportements de consommation, reflétant les transformations de notre société (notamment à la suite de la crise du Covid-19) ;

- KYC (know your customer) : détection des moments de vie (déménagement, mariage), de nouveaux profils clients (par ex. profil écologique) ;

- Risque de crédit : seuils personnalisés du taux d'endettement et du reste à vivre ;

- Service client : améliorer la gestion des dépenses selon ses habitudes, permettre de localiser chaque opération pour un contrôle toujours plus simple.

POURQUOI UNE TELLE ZONE DE FLOU ?

Principalement car les établissements bancaires n'ont pas le contrôle sur les informations contenues dans les libellés des transactions, et demeurent tributaires de la chambre de compensation. Ainsi les banques n'ont pas la main en cas de mise à jour du libellé d'une transaction. Par exemple, un libellé peut passer du nom complet de l'entreprise à un diminutif ou un sigle. De même, ils peuvent être différents entre deux terminaux de paiement d'un même magasin. Des difficultés qui complexifient toute automatisation de la catégorisation des opérations.

Une normalisation des transactions serait à envisager, ce qui amènerait une certaine harmonisation des libellés pour l'ensemble des établissements financiers et ainsi faciliter la lecture des opérations.

MÉTHODOLOGIE & RÉSULTATS

Dans ce cadre, l'Open Data constitue une piste à explorer dans le sens où la majorité des opérations carte présentent un libellé explicitement en lien avec le nom du commerçant : son nom juridique, ses noms d'usage, le tout couplé à une information géographique (ville, quartier, lieu-dit) totale ou partielle. Or les bases INSEE Sirene contiennent ce type d'informations (noms entreprises, information géographique).

L'intuition est donc de faire correspondre, pour chaque opération bancaire, une entreprise, un secteur d'activité et une localisation.

Dans cet usage, sont considérées uniquement les factures cartes (exclusion des prélèvements, virement, chèque, retrait, etc.) réalisées en France.

DONNÉES UTILISÉES

Nous utilisons un échantillon d'opérations bancaires issues de différentes banques (format .csv brutes).

Deux référentiels Open Data ont été utilisés dans le cadre de ce cas d'usage :

- Les données Sirene ;

- Les données Codes Officiels Géographique.

Les attributs qui la constituent : une catégorie d'opération, la date de l'opération, un nom (vraisemblablement composé d'un nom d'enseigne et d'une zone géographique), un identifiant du moyen de paiement, un numéro de compte, un code pays et un montant associé à une devise.

ANALYSE FONCTIONNELLE DE LA TRANSACTION

Le parsing de l'opération (action de structurer du texte) consiste à différencier et isoler l'ensemble des éléments constitutifs d'un libellé, ce qui revient à distinguer :

- Le nom de l'enseigne ;

- La localisation (ville, lieu-dit, dénomination, code département) ;

- Le montant ;

- Le numéro de carte ;

- Le code pays du lieu de transaction.

Bien que la méthode de parsing soit complexe et fasse appel à plusieurs étapes, nous prendrons l'exemple d'un cas simple :

FACTURE CARTE DU 300121 LEGO STORE PARIS CARTE 0000XXXXXXXX0000 FRA 19,99EUR

Un parsing fonctionnel du libellé complet de l'opération permet l'identification des différents attributs qui la constitue : une catégorie d'opération, la date de l'opération, un nom (vraisemblablement composé d'un nom d'enseigne et d'une zone géographique), un identifiant du moyen de paiement, un numéro de compte, un code pays et un montant associé à une devise.

L'intuition est donc de faire correspondre, pour chaque opération bancaire, une entreprise, un secteur d'activité et une localisation.

Dans cet usage, sont considérées uniquement les factures cartes (exclusion des prélèvements, virement, chèque, retrait, etc.) réalisées en France.

DATE DE RÉALISATION DE LA TRANSACTION	300121
NOM	LEGO STORE
LOCALISATION	PARIS
CARTE	0000XXXXXXXX0000
CODE PAYS	FRA
MONTANT	19,99
DEVISE	EUR

LE MATCH DU COUPLE <NOM, LOCALISATION>

Les bases Sirene présentent différents attributs permettant de caractériser une entreprise : sa dénomination légale, jusqu'à 3 dénominations usuelles, et encore davantage en incluant l'historique de cette entreprise. En plus de ces dénominations, les bases Sirene intègrent également la localisation exacte des entreprises.

Notre démarche s'appuie ensuite sur une succession d'algorithmes techniques, permettant de rapprocher le couple <nom, localisation> aux informations de ces bases Sirene. En cas de résultats multiples (plusieurs établissements semblent correspondre au couple <nom, localisation>), une succession de règles expertes et fonctionnelles sont alors appliquées, permettant ainsi d'associer la transaction à un établissement unique (SIRET unique).

Exemples de processus de déduplication des résultats :

- Ne répertorier que les établissements actifs à la date de transaction ;

- Priorisation des secteurs d'activité : les entreprises du secteur d'activité 56.10A (Restauration traditionnelle) ont une probabilité plus élevée d'être la réponse attendue contrairement aux sociétés avec un code NAF 94.20Z (Activités des syndicats de salariés).

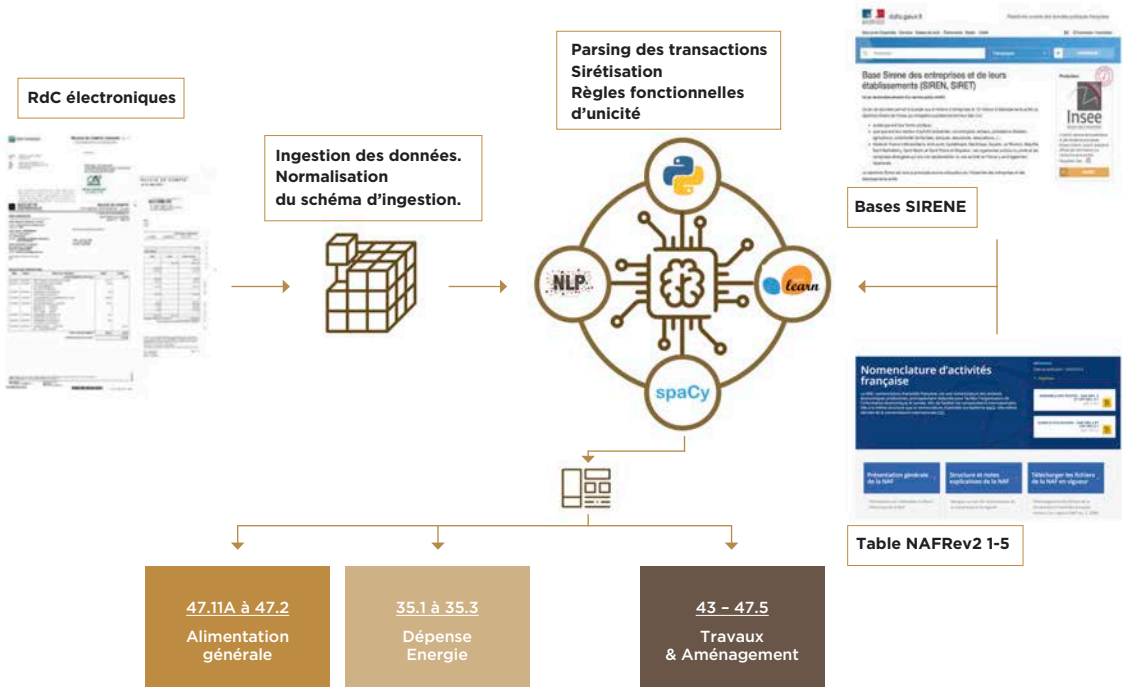
Sur notre exemple, l'application de l'algorithme permet d'obtenir les associations d'informations suivantes :

SIRET	75252617800096
DENOMINATION	LEGO STORE
VILLE	PARIS 1ER ARRONDISSEMENT
ADRESSE	101 RUE BERGER
CODE POSTAL	75001
CODE NAF	47.65Z (COMMERCE DE DÉTAIL DE JEUX ET JOUETS EN MAGASIN SPÉCIALISÉ)

Cette dépense est associée à la catégorie du code NAF « Commerce de détail de jeux et jouets en magasin spécialisé » là où la banque concernée par cette transaction l'a classifiée dans la catégorie « Enfant – autre ».

Autre exemple très parlant sur la plus-value apportée d'une telle labellisation : les dépenses réalisées au sein de l'enseigne Décathlon (grande distribution du sport) et Fitness Park (salle de sport) sont catégorisées dans « Sport » par la banque. Ici les codes NAF de Décathlon et Fitness Park correspondent respectivement à 47.64Z « Commerce de détail d'articles de sport en magasin spécialisé » et 93.13Z « Activités des centres de culture physique ». Un exemple qui montre une distinction entre achat d'articles sportifs et pratique sportive.

Le schéma suivant reprend en synthèse les étapes de construction du processus de « siretisation » et de catégorisation des opérations décrites précédemment.



RÉSULTAT

Sur l'ensemble des opérations, 81% d'entre elles ont pu être associées à un SIRET. Parmi ces opérations, le taux d'exactitude de l'association du SIRET est de 78%. C'est ainsi 61% des opérations qui sont correctement labellisées.

Quelle marge de progression et d'amélioration ? En tenant compte, en situation réelle dans un établissement bancaire, des volumes que représentent chaque couple <nom, localisation>, des rapprochements manuels ne sont pas à exclure pour traiter les cas atypiques et améliorer le taux de matching des opérations. En complément, les établissements bancaires disposent d'une information clef, le Merchant Category Code, caractérisant le secteur d'activité du commerçant. Une correspondance <nom, localisation, secteur d'activité> permettrait alors d'améliorer considérablement la précision des opérations de « siretisation ».

A travers ce cas d'usage, les Open data démontrent une nouvelle fois leurs pertinences et la diversité des besoins qu'elles peuvent servir, en étant cette fois-ci des inputs à l'enrichissement de la connaissance client. La fonctionnalité de rapprochement des transactions bancaires aux établissements étant suffisamment générique, elle peut ainsi être au cœur de nouvelles solutions, d'une catégorisation plus complète et localisée des dépenses visibles sur les espaces personnels clients, à de nouvelles offres de financement sur-mesure liées au profil de consommation des clients.

6.

CONCLUSION

Les Open Data offrent la possibilité d'accroître la diversité et la qualité de service et ce, quel que soit le domaine d'activité considéré. Leur principale force réside dans leur capacité à contextualiser la connaissance Client, en particulier sur le segment des Entreprises, Professionnels et TPE. Elles permettent d'accéder à un vaste ensemble d'informations et constituent donc un grand intérêt pour l'acquisition de la connaissance client, notamment pour clients récents ou Prospects. C'est un avantage concurrentiel indéniable pour les établissements lors de l'entrée en relation ou la détection d'opportunités de développement sur des segments de clientèle inexploités. Enfin, les Open Data peuvent s'imposer dans le cadre de la résolution de problématiques telles que le contrôle de la qualité des données, l'insuffisance de données ou encore l'insuffisance de profondeur d'historique, sans compter les perspectives pertinentes offertes en matière de feature engineering.

Comme le montre les cas d'usage développés, les Open Data permettent d'apporter des réponses concrètes et performantes à des problématiques métiers. Dans le cas de la détection de la cessation, il est possible d'envisager différents champs d'application tels que : la détection des risques de rupture d'approvisionnement liée à la cessation dans le domaine des Achat/Supply Chain, la détection d'opportunités d'investissement, la reprise de fonds de commerce ou d'activité, l'attribution d'aides au développement pour les secteurs à risque ou encore l'utilisation des modèles de cessation en complément des modèles internes de probabilité de défaut pour les établissements bancaires.

Toutefois, l'utilisation des Open Data nécessite le déploiement d'un effort de recherche et leur recours dans le cadre de processus opérationnels devra passer par une phase d'acculturation et par la démonstration de leur capacité à résoudre des problématiques métiers que les données internes seules ne permettent pas d'instruire. Dès lors, les établissements devraient entamer un processus d'exploration et catégoriser les Open Data utiles par domaine d'application, les intégrer à leur système d'information afin d'en permettre l'exploitation systématique.

Quelles que soient les contraintes opérationnelles et organisationnelles qui restreignent l'utilisation des Open Data, leur recours s'avère incontournable et constitue un enjeu majeur pour les établissements dans le cadre de la gestion des risques climatiques.

DONNER DU FUTUR AU TALENT

Fondé en 2008, Square est un groupe de conseil en stratégie et organisation qui réunit 7 cabinets en France, Belgique et Luxembourg. Adway, Circle, Flow&Co, Initio Belgique, Initio Luxembourg, Tallis, Vertuo, Viatys sont des cabinets de conseil spécialisés par métier, secteur d'activité ou niveau d'intervention.

Cette organisation, unique et spécifique, favorise la proximité, l'engagement, l'agilité et l'expertise au sein de chaque cabinet. La complémentarité des cabinets permet à Square d'adresser, avec plus de 700 consultants, les projets les plus complexes de ses clients. Square conseille ses clients en mettant à leur disposition ses expertises sur 9 domaines phares.

DATA
Square élabore des stratégies Data et assure leurs déclinaisons opérationnelles à travers la conduite de projets de Data Management, Data Analyse et Data Science. Notre approche experte et pragmatique vise à valoriser et sécuriser le patrimoine de données des entreprises.

DIGITAL
Square accompagne ses clients dans l'élaboration de leur stratégie digitale, la conception et la mise en œuvre de nouveaux parcours digitaux pour leurs clients ou leurs collaborateurs, ainsi que dans l'ensemble des chantiers d'acculturation interne et d'accompagnement aux nouvelles méthodes de conception.

INNOVATION
Square accompagne ses clients dans la transformation de leur dynamique d'innovation. Nos consultants, par leur approche sur-mesure, aident à concevoir, industrialiser et gouverner l'innovation pour assurer la croissance durable des entreprises et leur transformation en entité socialement et écologiquement responsable.

MARKETING
Square accompagne ses clients sur l'ensemble du spectre marketing : marketing stratégique, marketing relationnel, marketing de l'offre, communication, tarification, satisfaction clients. Nos expertises initialement centrées sur les secteurs de la banque et de l'assurance, s'adressent désormais à l'ensemble des industries ou services B2C.

PEOPLE & CHANGE
Square aide ses clients à acquérir, fédérer et développer le capital humain de leur organisation. Afin de créer davantage d'engagement au sein des équipes, nos interventions portent principalement sur l'adaptation des méthodes de travail aux changements opérationnels et culturels, l'efficacité des directions des ressources humaines et le développement des compétences.

RISK & FINANCE
Square prend en charge le pilotage des programmes de maîtrise des risques financiers et non financiers, ainsi que la transformation des fonctions Risque et Finance face à l'évolution des dispositifs prudentiels et à l'irruption des problématiques liées à la maîtrise de la donnée.

REGULATORY & COMPLIANCE
Square conseille ses clients dans le déploiement des nouvelles réglementations, ainsi que dans l'optimisation et le renforcement des dispositifs de contrôle. Ce domaine d'excellence s'appuie sur une communauté d'experts de 130 consultants qui, outre ses missions auprès des clients, conduit d'importants travaux d'investigation et de publication.

RSE ET FINANCE DURABLE
Square accompagne ses clients dans leur transformation vers un modèle plus responsable.Cet accompagnement porte sur la définition stratégique de l'ambition RSE, la transformation des business models, les travaux de mise en conformité tant dans leur déclinaison réglementaire que dans leur déclinaison Data Management et Data Science. Square accompagne également ses clients dans leurs chantiers d'accompagnement humain et culturel de leur politique RSE.

SUPPLY-CHAIN
Square assure l'excellence opérationnelle de la logistique, des achats aux derniers kilomètres, avec des parcours clients différenciants. Nos experts conçoivent des solutions omnicanales mettant en œuvre les meilleures pratiques des systèmes d'informations, de la mécanisation à la robotisation.

Ce nouveau Focus Square préparé par les consultants du Domaine d'Excellence Data revient sur une opportunité encore peu exploitée par les institutions financières : l'open data.

Concentrées sur l'organisation et l'exploitation de leur propre patrimoine data, les organisations peuvent avoir tendance à ne pas prendre en compte les gains possibles par le recours aux données publiques lorsqu'elles déclenchent des initiatives reposant sur la donnée.

Dans ce document, nous proposons ainsi de revenir sur la montée en puissance de la donnée publique et de partager quelques cas concrets de valeur métier apportée par la donnée publique dans un projet Data.

CONTACT

ADRIEN AUBERT
Associate partner
adrien.aubert@vertuoconseil.com

Square

Crédits photos - Unsplash : Laureen Missaire (p. 1), Erol Ahmed (p.2), Artem Zhukov (p.6), Zanr Lee (p.22).

Square

DONNER DU FUTUR AU TALENT

square-management.com