Master d'Ingénierie Statistique et Financière

Prévision des entrées en premier impayé :

Amélioration de la prévision du coût du risque

Présenté par Hugo BREHIER.

Mémoire de M2

Préparé sous la direction de Olivier BOS, Maître de conférences, Paris II Panthéon-Assas.

Soutenu en Septembre 2019

À l'Université Paris II Panthéon-Assas.

Année universitaire 2018-2019

Master d'Ingénierie Statistique et Financière

Prévision des entrées en premier impayé :

Amélioration de la prévision du coût du risque

Présenté par Hugo BREHIER.

Mémoire de M2

Préparé sous la direction de Margerie AUTIN, Chargée d'études.

Soutenu en Septembre 2019

À l'Université Paris II Panthéon-Assas.

Ce mémoire a été préparé dans le cadre d'un apprentissage au sein de :

Crédit Agricole Consumer Finance

17 Rue Victor Basch, 91300 Massy

De septembre 2018 à septembre 2019.

Tout d’abord, je remercie ma maître d’apprentissage, Margerie AUTIN, pour sa bienveillance.

Je remercie aussi mon équipe à CACF de m’avoir bien accueilli.

Egalement, Monsieur le Professeur Georges BRESSON pour son aide précieuse.

Finalement, je remercie toutes les personnes qui m’ont permis d’accomplir ce travail d’une année.

La prévision de données temporelles à des horizons de 6 mois, 12 mois et 18 mois est toujours un sujet de discussion brûlant. A des méthodes statistiques traditionnelles s'ajoutent des méthodes plus récentes d'intelligence artificielle. Le sujet de ce rapport sont les entrées en impayé des crédits à la consommation de Crédit Agricole Consumer Finance (CACF). Leur prévision doit permettre d'établir le budget de l'année en cours pour la *Business Unit* France.

Dans ce rapport, je décris toutes les étapes d'une étude statistique, avec pour aboutissement la prévision effective des entrées en impayé, pour les 18 prochains mois.

Mots-clefs : prévision, VECM, ARIMA, RNN, impayés

Sommaire

Introduction

Parcours universitaire

Mon début dans les études supérieures s'est déroulé à l'Université de Versailles Saint-Quentin-en-Yvelines, au cours d'une licence d'Economie.

Cela m'a permis d'appréhender l'univers bancaire et financier, la macroéconomie et la microéconomie. J'ai découvert, au fil de ces années, un intérêt spécial pour les statistiques.

La poursuite de mes études a continué naturellement dans le master d'Ingénierie Statistique et Financière de l'Université Paris II Panthéon-Assas.

Au cours de la première année, qui s'est déroulée en alternance, j'ai découvert l'univers des risques bancaire au sein de La Banque Postale. Chargé d'automatisation de reportings, j'ai pu acquérir beaucoup de notions sur la solvabilité, la liquidité, les demandes règlementaires de Bâle, de la BCE et de l'ACPR.

J'ai alors décidé qu'il était temps pour moi de mettre en pratique mes connaissances statistiques. Pour ma dernière et actuelle année de master, j'ai ainsi rejoint le Groupe Crédit Agricole au sein de l'entité Crédit Agricole Consumer Finance, en tant que chargé d'études statistiques.

Présentation de l'entreprise

Crédit Agricole Consumer Finance (CACF) est l'entité du Groupe Crédit Agricole offrant des solutions de prêt à la consommation : prêt personnel, crédit renouvelable, financement automobile, distribution, etc.

Crédit Agricole Consumer Finance est né de la fusion entre Sofinco et Finaref en avril 2010. Le groupe Crédit Agricole en est le seul actionnaire. Depuis 2013, Finaref a été renommé Sofinco, qui devient donc la marque de CACF en France. C'est en 2018 le 2ème acteur du crédit à la consommation en France.

En tant qu'entité du groupe Crédit Agricole, CACF coopère aussi avec les autres parties du groupe (LCL, BForBank, caisses régionales) en développant des offres en France de crédit à la consommation.

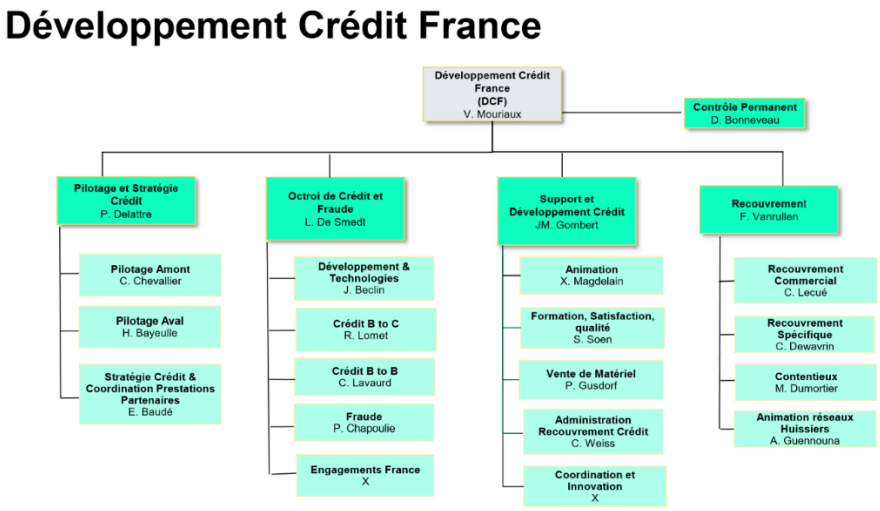
L'entreprise est présente dans 17 pays d'Europe, en Chine et au Maroc et compte 10 000 collaborateurs. Cela en fait le troisième plus grand acteur du crédit à la consommation en Europe avec 82 milliards d'euros d'encours géré et 2000 millions d'euros de Produit Net Bancaire.

Graphique 1 : Activité Européenne de CACF

J'ai intégré CACF au sein de la direction Développement Crédit France (DCF). En tant que support à la production de crédit en France, la direction a plusieurs missions attitrées :

- Fluidifier et accélérer le parcours d'octroi sans dégrader le risque d'impayé et de fraude
- Améliorer la sélection et la qualification des clients
- Simplifier et harmoniser les règles et processus d'acceptation
- Moderniser les outils d'octroi et de fraude
- Prendre en charge l'étude, l'acceptation, la production et la gestion des dossiers d'engagement entreprise

C'est dans le cadre de cette dernière mission que se place mon année d'alternance, plus particulièrement sur la mission d'étude au sein du service Pilotage et Stratégie Crédit (PSC) et du Pilotage Amont.



Graphique 2 : Organigramme de DCF

PSC a plusieurs missions de support au sein de DCF :

- Prévoir, suivre et analyser les évolutions du risque de crédit en France, et accompagner les évolutions des processus amont et aval à travers le pilotage des stratégies
- Animer la ligne métier Crédit France à travers la diffusion des tableaux de bord Crédit et la coordination de la gouvernance Crédit France
- Assurer le pilotage du risque des périmètres pour compte de tiers et son évolution dans le cadre du développement du servicing
- Développer la prestation pour compte de tiers à travers la structuration des relations et la fidélisation des partenaires actuels, la communication et le développement de notre offre de servicing pour le compte du groupe Crédit Agricole, le développement de la conquête externe permettant d'offrir nos prestations internes en matière de recouvrement

Problématique et plan d'action

Ma mission d'alternant se situe donc dans le Pilotage Amont des risques d'impayés, l'entreprise se devant d'envisager ses futurs risques de crédit liés à son activité.

Les indicateurs utilisés sont le nombre de crédits auparavant sains entrant en défaut (pour la première fois ou non) et leur montant. Ceux-ci sont des éléments du calcul du coût du risque calculé lors du budget annuel d'avril, qui synthétise le poids des créances dans le bilan des banques françaises.

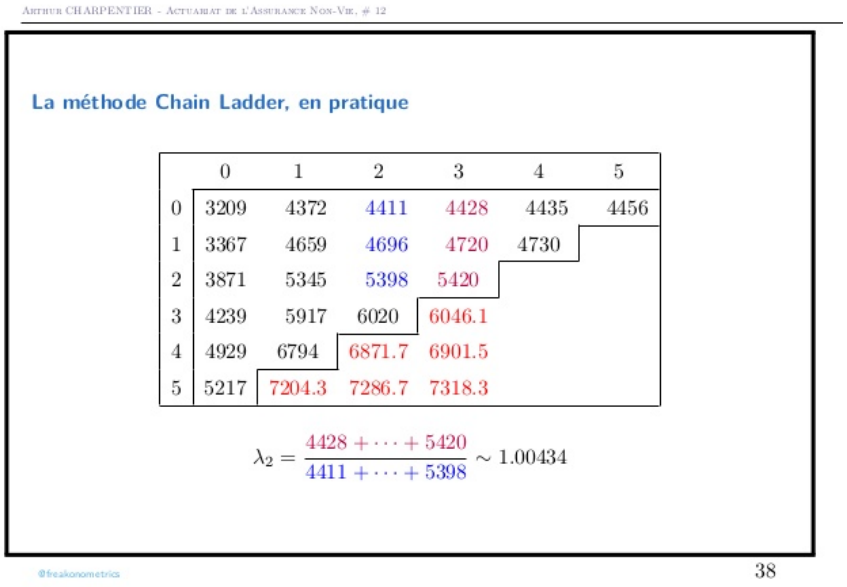
Ma tâche principale au cours de cette année a été d'apporter mes connaissances statistiques à la construction d'un modèle de prévision des entrées en impayés des crédits à la consommation de CACF en France.

Dans l'élaboration de ces prévisions, j'ai coopéré avec une autre étudiante de PolyTech Lille en Génie Informatique et Statistique en alternance dans mon service, mais sur un autre site. Celle-ci était déjà

présente l'année dernière dans ce service, année durant laquelle elle s'est penchée sur des modèles ARIMA.

De plus, ma tutrice d'entreprise m'a guidé tout au long de l'année par son savoir métier, à l'aide de son expérience longue de plusieurs années au sein de l'entreprise.

La méthode précédente à la mienne de prévision des nombres d'impayés a été pensée par ma tutrice. Elle consiste en un allongement des taux de croissance précédemment observés dans chaque génération de production. Cela s'apparente ainsi à la méthode actuarielle bien connue, dénommée *Chain Ladder*[1].



Graphique 3 : Méthode dite *Chain Ladder*

Il faut un certain savoir métier pour ajuster certains taux de croissance incohérents, notamment après des allers-retours avec le service de pilotage commercial quant à des hypothèses de production de crédit.

Cette méthode ne permet aussi malheureusement pas de fournir des explications quant à ses prévisions. Dans ce but de fournir des explications, il a été décidé en amont de mon arrivée qu'incorporer des données macroéconomiques dans la modélisation serait utile.

Une étude a ainsi été commandée au cabinet QuantMetry, qui accompagne les entreprises dans leurs projets *data* (Alstom, SNCF, Microsoft)[2]. Celle-ci s'est conclue par un modèle constitué de régressions linéaires sur les *lags* de l'historique des nombres d'impayés, sans prise en compte de données macroéconomiques.

[1] https://freakonometrics.hypotheses.org/tag/chain-ladder
[2] https://business.lesechos.fr/entrepreneurs/marketing-vente/0301760163621-organiser-un-salon-un-booster-commercial-pour-quantmetry-322589.php

Mon arrivée coïncide donc avec la demande renouvelée de fournir un modèle statistique incorporant des données externes à l'entreprise. L'étude que j'ai entreprise a été constituée de :

- Exploration des données externes
- Recherche des modèles potentiels
- Validation des modèles
- Lancement des prévisions pour le budget en cours
- Présentation au Comité Exécutif de CACF du projet

Dans une première étape de validation, les données s'organisent comme ceci :

- Janvier 2013 à juin 2017 : échantillon d'estimation
- Juin 2017 à décembre 2018 : échantillon de validation

Dans une deuxième étape de prévision réelle :

- Janvier 2013 à avril 2019 : échantillon d'estimation
- Avril 2019 à décembre 2020 : période future prédite

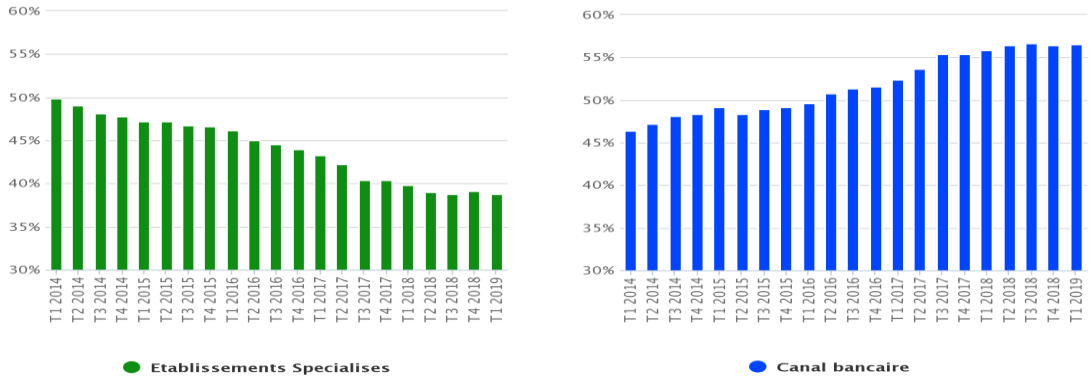
Les données sont également séparées par segment d'activité.

Première partie

Contexte global

Crédit Agricole Consumer Finance est issu de la fusion de Sofinco et Finaref en 2010. Aujourd'hui, c'est un des 3 acteurs principaux du crédit à la consommation en Europe et en France. Le premier acteur en Europe et en France étant BNP Paribas Personal Finance, crée en 2007 par la fusion de Cétélem et l'UCB. Le troisième acteur en France est Cofidis, racheté en 2008 par le Crédit Mutuel.

Les établissements de financement spécialisés sont les leaders historiques du marché français, en baisse de nos jours, pour atteindre un peu moins de 40% des crédits à la consommation distribués.



Graphiques 4-5 : Encours Français de crédit à la consommation des établissements spécialisés et des établissements généralistes[3]

Les établissements bancaires généralistes ont pris une part de plus en plus importante dans la distribution de crédits à la consommation en France. Crédit Agricole Consumer Finance fait partie de l'écosystème du groupe Crédit Agricole. En France, cela lui offre divers débouchés pour commercialiser ses produits. Une partie de son activité est dédiée à l'offre de produits directement dans ses agences Sofinco, une autre indirectement dans les agences des caisses régionales du groupe, aux agences LCL ou encore sur BForBank.

Les sociétés automobiles ou de distribution comptent parfois sur une société captive pour proposer des solutions de financement et d'assurance à leur client, au lieu d'établir un partenariat avec un établissement de financement spécialisé.

En France, les crédits à la consommation connaissent une phase de croissance importante, avec plus de 5% de croissance au dernier semestre 2018 et au premier semestre 2019. Le total des crédits à la consommation accordés aux particuliers est de 181 milliards d'euros. Cela est soutenu principalement par les prêts amortissables et les crédits-bails (secteur automobile).

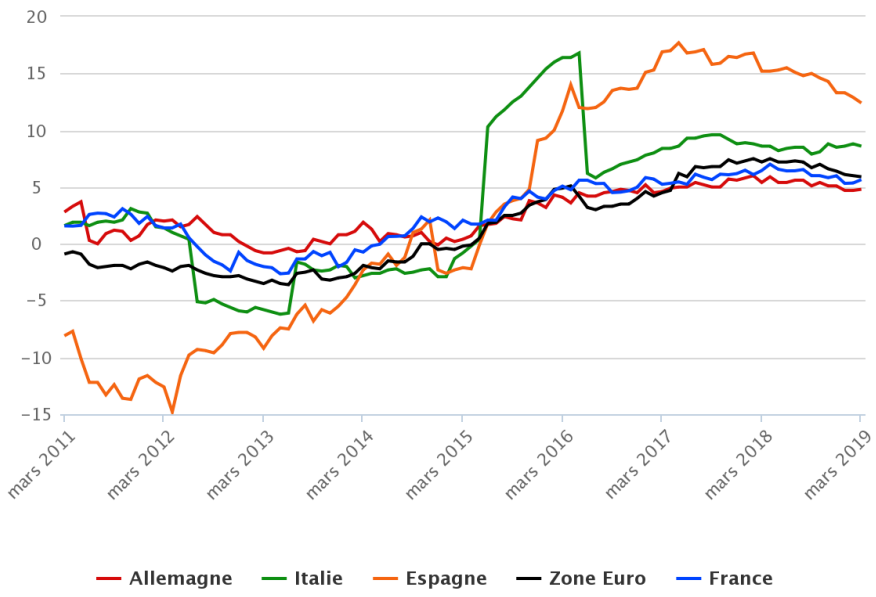
[3] https://www.banque-france.fr/statistiques/credit/credit/credits-la-consommation

France, milliards d'euros, CVS

	2017		2018				2019	Taux de croissance annuel	
	Sept.	Déc.	Mars	Juin	Sept.	Déc.	Mars	Déc. 2018	Mars 2019
Total crédit à la consommation aux particuliers	**166,1**	**169,0**	**170,8**	**173,6**	**175,6**	**179,1**	**180,9**	**6,1%**	**5,6%**
dont Prêts amortissables y compris créances titrisées	115,9	117,7	118,3	119,9	120,9	123,3	124,7	4,6%	5,1%
Comptes ordinaires débiteurs	8,1	8,0	8,2	8,1	8,2	8,3	8,2	3,6%	0%
Crédits renouvelables	18,7	18,7	18,8	18,8	19,1	19,1	19,1	2,1%	1,6%
Crédits-bails	11,3	12,3	12,9	13,8	14,5	15,0	15,5	18,0%	16,8%

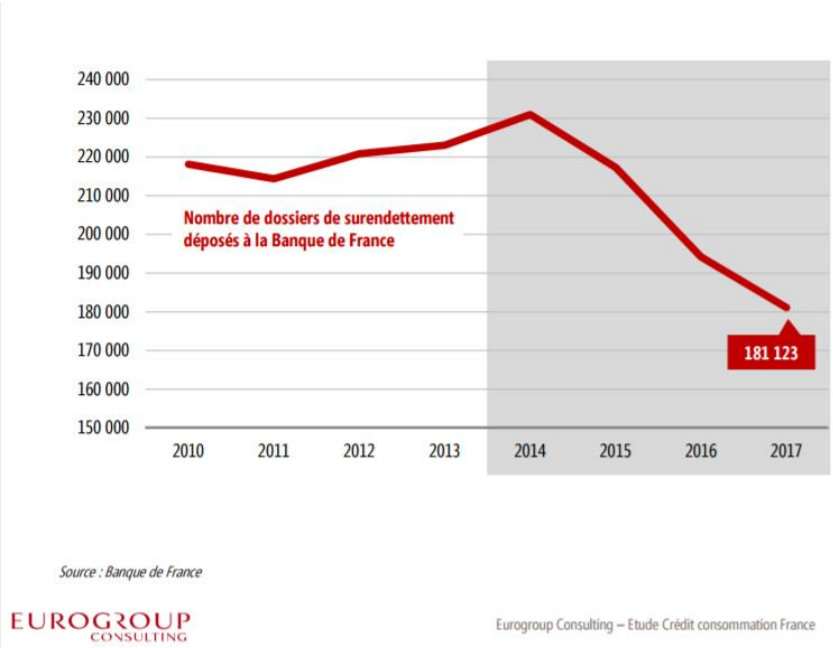
Table 1 : Evolution des crédits à la consommation en France[3]

En Europe, la tendance est à une croissance légèrement infléchie depuis 2018. Cela est dû à la politique de taux directeur bas de la Banque Centrale Européenne, qui encourage les particuliers à emprunter pour relancer l'activité.



Graphique 6 : Taux de croissance des crédits à la consommation en Europe[3]

De plus, le nombre de dossier en surendettement est en baisse depuis 2016. Or, le secteur est réputé pour un risque de crédit élevé et des taux d'intérêt élevés, toutefois à la baisse depuis 2014. Cela va de pair avec la baisse relative d'importance des crédits renouvelables par rapport aux produits amortissables et au crédit-bail (table 1).



Graphique 7 : Nombre de dossiers de surendettement[4]

Cependant, le début d'année 2019 dénote des faiblesses du secteur. Structurellement, les acteurs du crédit à la consommation n'ont pas accès à l'épargne des particuliers à travers les comptes courants bancaires. L'univers européen de taux bas (le taux moyen des crédits à la consommation est descendu à 3.5% en 2019) est dur à supporter à long-terme, en plus des exigences de fonds propres bâloises et des lois régulant le prêt renouvelable, principal instigateur du surendettement.

Graphique 8 : Lois sur le crédit renouvelable[4]

[4] https://www.eurogroupconsulting.com/sites/eurogroupconsulting.fr/files/document_pdf/eurogroup_consulting_etude_credit_consommation_vf.pdf

C'est un “sujet structurel de profitabilité du crédit à la consommation” dont parle Jean-Marie Bellafiore, directeur général délégué de BNP Paribas Personal Finance[5]. L'apparition des gilets jaunes, puis les mesures sociales prises par le gouvernement rendent l’année très incertaine.

Ecosystème interne de DCF

CACF est doté d’un système de support à la commercialisation de ses offres en France, qui se retrouve en partie au sein de DCF.

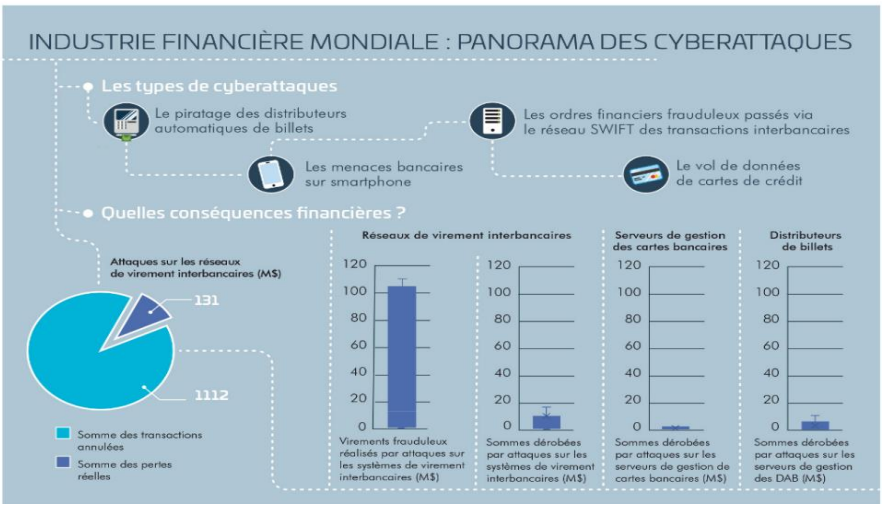
L’octroi de crédit est appuyé sur un système de scores. Ce système est basé sur les caractéristiques individuelles des clients et donne un avis favorable ou non à l’octroi d’un crédit. Cela permet aux conseillers en agence d’être épaulé par un avis technique avant de décider de l’octroi d’un crédit.

Ce système est actuellement constitué d’un ensemble de systèmes experts. Chaque score doit être contrôlé. Une partie des équipes de la direction s’en occupe. Il faut ainsi s’assurer du poids des variables, de leur cohérence avec le bon sens et de leur accord avec les lois éthiques. Un client doit pouvoir obtenir une explication valable sur un refus de crédit.

Le recouvrement des impayés est une autre tâche effectuée au sein de la direction. Il s’agit, comme le suggère le nom, de permettre à la banque d’atténuer les effets d’un défaut de paiement. C’est cette procédure qui mène *in fine* à la création d’un dossier de surendettement.

La gestion de la fraude est importante. Au-delà de la fraude traditionnelle, les systèmes informatiques bancaires sont des cibles très prisées des cybercriminels[6]. L’intelligence artificielle pour détecter la fraude et la cyberdéfense des systèmes sont des sujets récents critiques pour la santé des entreprises.

SEKOIA THALES



Graphique 9 : Les vecteurs de cybercriminalité[7]

[5] https://www.lesechos.fr/finance-marches/banque-assurances/lannee-2019-a-mal-commence-pour-le-marche-du-credit-a-la-consommation-1001973

[6] https://www.lesechos.fr/finance-marches/banque-assurances/banques-les-nouvelles-techniques-des-cyber-braqueurs-959488

[7] https://www.thalesgroup.com/fr/marches-specifiques/systemes-dinformation-critiques-et-cybersecurite/news/thales-et-sekoia

Finalement, les risques de crédit doivent être suivis. C'est le rôle du pilotage des risques, qui s'occupe de suivre des indicateurs, notamment certains dits "à chaud", permettant d'établir des stratégies efficaces d'octroi, de recouvrement ainsi que de lutte contre la fraude.

Problématique

Au sein du Pilotage des Risques en amont, la question de la prévision de certains indicateurs se pose donc. Le nombre d'impayés et leur montant par segment d'activité rentrent dans le calcul du coût du risque et donc dans l'établissement du budget de la direction (DCF) pour l'année en cours. Cette opération a généralement lieu au début du printemps. Il faut ainsi effectuer une prévision du nombre et du montant d'impayés pour pouvoir effectuer ce budget.

L'ancienne méthode de prévision des impayés devait être refondue en une méthode robuste, rigoureuse, spécifique et rapide :

- Robuste, car il faut une méthode qui soit précise à court-terme, c'est-à-dire à moins d'un semestre, et à long-terme, c'est-à-dire à 18 mois.
- Rigoureuse car il faut ajouter une rigueur statistique à des méthodes jusque-là *ad-hoc*. Un système de validation pour la sélection du modèle et des données doit être mis en place.
- Spécifique car il faut incorporer des données externes à la modélisation avec comme objectif l'inspection du modèle pour expliquer les tendances futures trouvées.
- Rapide car il y a plus de 20 segments à traiter, et les délais du processus de budget ne permettent pas de prendre un temps infini.

A noter que les prévisions sont faites mensuellement mais évaluées semestriellement.

Méthodologie de recherche

Les données utilisées sont donc de deux origines : internes et externes.

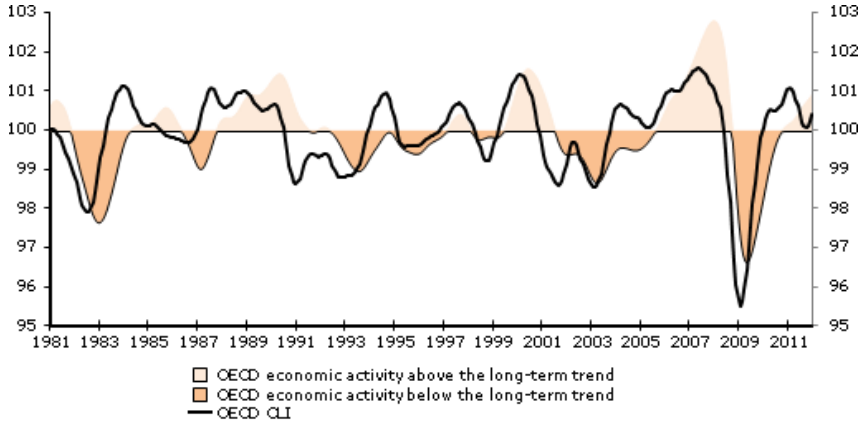
Tout d'abord, il y a un historique interne des impayés sous un format agrégé pour économiser de la place sur le serveur d'entreprise. J'ai demandé s'il était possible d'obtenir la base au niveau dossier, c'est à dire désagrégé, mais ce n'a pas pu être le cas. Cela aurait pris trop de place sur le serveur.

Ainsi, nous sommes en présence de données temporelles caractérisées par :

- Un mois d'apparition
- Un segment d'activité
- Un nombre et un montant d'impayés

A cet historique, il faut adjoindre des données macroéconomiques. La question se pose desquelles choisir et par quelle méthode. Dans la littérature économique, il existe deux types d'indicateurs.

Les *Leading Indicators*, en avance sur l'économie dans son ensemble. Ils permettent d'effectuer des prédictions sur la dynamique future de l'économie. L'OCDE a par exemple conçu un indicateur, le Composite Leading Indicator (CLI), qui doit permettre d'anticiper les cycles économiques. Le graphique ci-dessous illustre l'intérêt d'un tel indicateur pour la prévision économique.



Graphique 10 : Le CLI de l'OCDE[8]

Les *laggings Indicators* sont des indicateurs en retard sur l'activité économique. C'est par exemple le cas du taux de chômage.

Les données externes choisies sont issues de l'Open Data, elles sont donc gratuites et en ligne. Les données ci-dessous sont celles contenues dans l'étude de QuantMetry:

- Euribor à 1,3,6 et 12 mois (Banque de France)
- Nombre d'immatriculations neuves et d'occasion (CCFA)
- Baltic Dry Index (Investing.com)
- Taux d'endettement des ménages en France (Banque de France)
- Encours total de crédit à la consommation en France (Banque de France)
- Confiance des ménages français (INSEE)
- Indice de référence des loyers (INSEE)

D'autres variables de l'études n'étaient pas utiles car mal documentées et donc sans possibilité d'être mises à jour. J'ai donc ajouté des données similaires :

- Consumer Confidence Index (OCDE)
- Taux de croissance du PIB français (INSEE)
- Taux d'inflation sous-jacente (INSEE)
- Taux de chômage (INSEE)

Les données internes sont mensuelles. Comme les données externes n'ont pas toutes cette fréquence, j'ai dû les adapter à une fréquence mensuelle. Pour les Euribor, journaliers, j'ai choisi la dernière valeur du mois. Pour plusieurs données, très agrégées, j'ai répliqué la valeur trimestrielle sur chacun des trois mois.

Vient également la question de la saisonnalité des variables. Les données internes sont brutes alors que certaines données des institutions publiques sont CVS, Corrigées des Variations Saisonnières[9]. On décompose usuellement une série temporelle en [10]:

[8] https://www.oecd.org/sdd/compositeleadingindicatorsclifrequentlyaskedquestionsfaqs.htm#1
[9] https://www.insee.fr/fr/metadonnees/definition/c1599
[10] https://www.math.univ-toulouse.fr/~lagnoux/Poly_SC.pdf

- Tendance, le comportement moyen de la série
- Saisonnalité, le comportement périodique, *i.e.* à intervalles réguliers, de la série
- Résidus de la série

Je n'ai pas corrigé les données internes de leur composante saisonnière, que je ne n'ai pas considéré comme impactant. Il existe également plus en aval des méthodes de prise en compte de la saisonnalité.

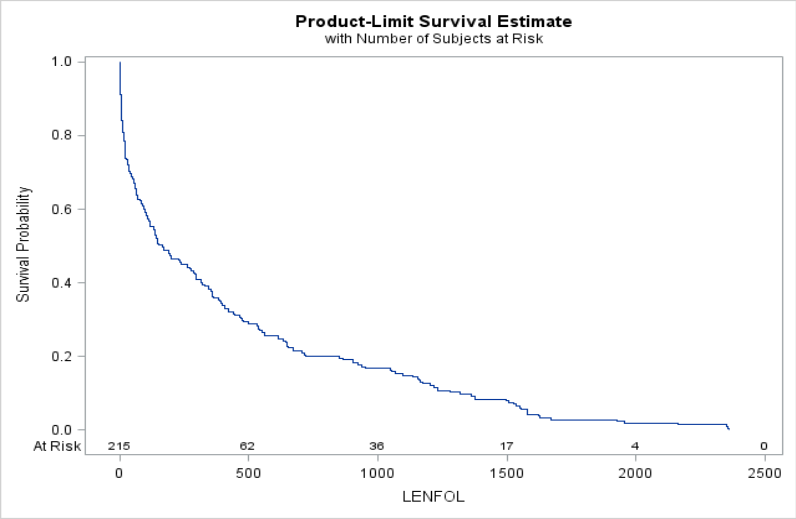
En rétrospective, il aurait peut-être été utile de faire plus attention à ces étapes de préparation. J'aurais en effet pu effectuer une interpolation des séries trimestrielles en séries mensuelles, une moyenne mobile de 30 jours pour l'Euribor et corriger les séries internes de leur saisonnalité en amont.

Une fois les données obtenues, il faut choisir un type de modèle.

Au début de mon alternance, j'ai considéré l'analyse de survie, pour essayer d'exploiter une hypothétique base de données au niveau dossier. L'analyse de survie permet de modéliser le temps écoulé avant la survenance d'un évènement, ici l'entrée en impayés. La fonction de survie est définie par :

$$S(t) = P(T > t)$$

Soit la probabilité que la variable aléatoire T subisse l'évènement après l'écoulement d'une période de durée t.



Graphique 11 : Fonction de survie sur des données d'accidents cardiaques[11]

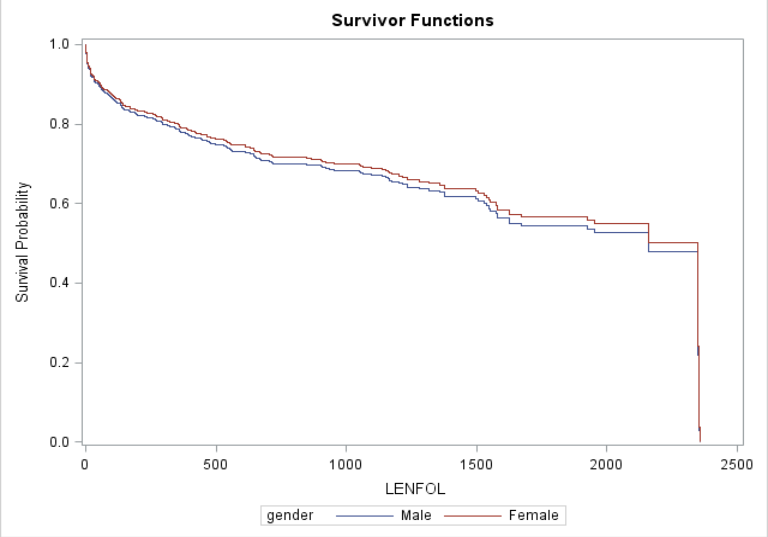
[11] https://stats.idre.ucla.edu/sas/seminars/sas-survival/

On définit également une fonction d'aléa, dénotant du risque instantané d'apparition de l'évènement à une date donnée t sachant qu'il n'est pas encore arrivé :

$$U(t) = f(t)/S(t)$$

Où *f(t)* est la densité de probabilité d'un évènement futur.

En créant des groupes homogènes de dossiers, *i.e.* en segmentant les dossiers par des covariables judicieuses, on peut estimer plusieurs fonctions de survie sur l'historique interne. En utilisant les fonctions estimées, on peut calculer pour chaque mois futur le nombre de dossiers qui entre en impayé, en supposant que les dossiers survivants suivront le même processus qu'estimé sur l'historique de leur groupe respectif.[12]



Graphique 12 : Fonctions de survie pour deux groupes, hommes et femmes

Je n'ai pas pu mettre en place une telle méthode, le responsable de l'équipe ne trouvant pas l'idée judicieuse pour des raisons pratiques d'extraction de données.

Je suis alors revenu aux méthodes de séries temporelles vues en première année de Master avec Monsieur le Professeur Georges Bresson[13]. C'est lui-même qui m'a mis sur la piste d'un modèle à correction d'erreur multivarié, un VECM (*Vector Error-Correcting Model*) alors que j'étudiais les modèles ARIMAX (*AutoRegressive Integrated Moving Average model with eXogeneous variables*).

Le modèle ARIMAX est une extension des célèbres modèles ARMA, avec incorporation de variables exogènes.

[12] https://www.ajol.info/index.php/orion/article/download/111549/101328
[13] http://bresson.u-paris2.fr/doku.php?id=start

Un modèle ARMA classique est défini par[14] :

$$y(t) = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \theta_1 z_{t-1} - \cdots - \theta_p z_{t-p} + z_t$$

Où z est un processus de bruit blanc.

Cela mêle processus autorégressif (AR) et processus de moyenne mobile (MA), qui doivent pouvoir approximer n'importe quel processus stationnaire, selon le théorème de décomposition de Wold[15].

C'est un modèle qui s'applique donc à des données stationnaires (faibles), c'est à dire dont la moyenne, la variance et l'autocorrélation sont stables par propriété de distribution inconditionnelle fixe dans le temps[16]. Il faut donc s'assurer de cette propriété, par intégration de la série. Ce que l'on fait en prenant la série différenciée, selon la méthodologie de Box et Jenkins[17].

En ajoutant des données exogènes, on obtient l'ARIMAX suivant :

$$y(t) = \beta X_t + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \theta_1 z_{t-1} - \cdots - \theta_p z_{t-p} + z_t$$

Mais le coefficient β n'a pas l'interprétation usuelle qu'il peut avoir dans une régression linéaire classique. En effet, l'effet β est conditionnel aux effets des valeurs passées de y, ce qui n'est pas instinctif. On préfère alors le modèle de régression avec erreurs ARMA :

$$y(t) = \beta X_t + n_t$$

$$n(t) = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \theta_1 z_{t-1} - \cdots - \theta_p z_{t-p} + z_t$$

On peut étendre la régression à plusieurs régresseurs, que ce soit par des lags différents ou de nouvelles séries d'input, ce qui donne encore un nom différent au modèle : une régression dynamique[18] (*dynamic regression model* ou *distributed lag model*).

Le VECM est quant à lui un modèle multivarié à correction d'erreur, qui permet d'utiliser des séries non-stationnaires à travers la cointégration[19]. Il modélise un système de séries temporelles interdépendantes en équilibre à long-terme (la relation de cointégration), subissant des déséquilibres à court-terme (les relations différenciées) et des chocs innovants[20] :

$$\Delta X_t = \Pi X_{t-1} + \sum_{i=1}^{p-1} \Phi_j^* \Delta X_{t-i} + \varepsilon_t$$

$$Où \ \Phi_j^* = -\sum_{i=j+1}^{p} \Phi_i \ , \qquad j = 1, \dots, p-1$$

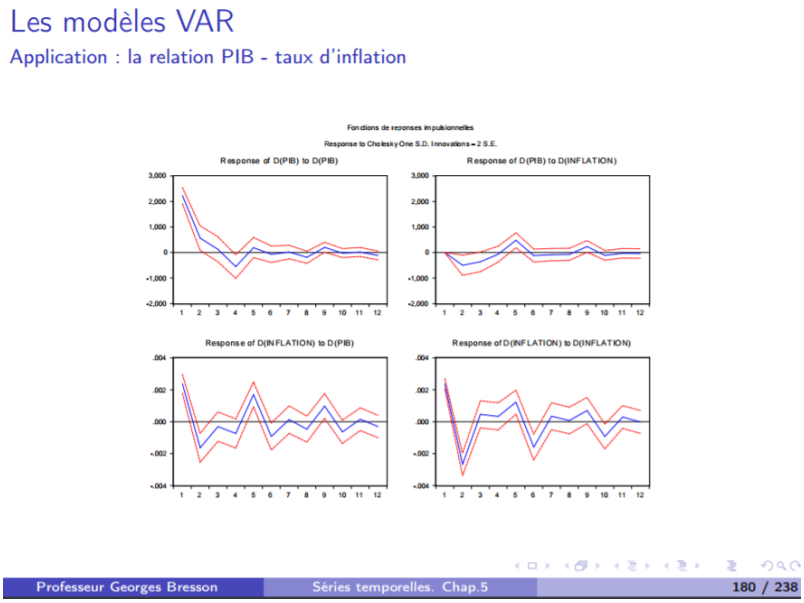
$$\text{et } \Pi = -(\mathrm{I} - \Phi_1 - \cdots - \Phi_p)$$

[14] https://robjhyndman.com/hyndsight/arimax/
[15] http://mayoral.iae-csic.org/timeseries2019/handout2_arma.pdf
[16] Gagniuc, Paul A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. USA, NJ: John Wiley & Sons. pp. 1–256. ISBN 978-1-119-38755-8.
[17] George Edward Pelham Box and Gwilym Jenkins. 1990. Time Series Analysis, Forecasting and Control. Holden-Day, Inc., San Francisco, CA, USA.
[18] http://www-personal.umich.edu/~franzese/DeBoefKeele.2008.TakingTimeSeriously.pdf
[19] Robert F. Engle and C. W. J. Granger, Co-Integration and Error Correction: Representation, Estimation, and Testing, Econometrica, Vol. 55, No. 2 (Mar., 1987), pp. 251-276

[20] http://statmath.wu.ac.at/~hauser/LVs/FinEtricsQF/FEtrics_Chp4.pdf

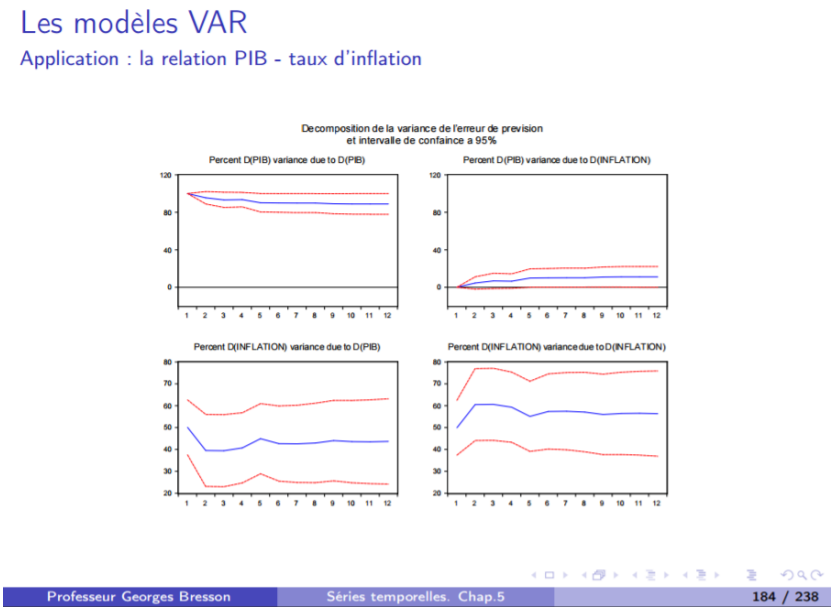
Cela lui confère ainsi un pouvoir explicatif intéressant, les coefficients ayant une signification réelle. On peut aussi utiliser, à des fins explicatives, la décomposition de la variance des erreurs de prédiction et les fonctions de réponses impulsionnelles.

Les fonctions de réponses impulsionnelles donnent la réponse prévue d'une variable à un choc innovateur dans une autre variable.



Graphique 13 : Fonctions de réponses impulsionnelles

La décomposition de la variance des erreurs de prédiction permet de savoir quelle variable impacte les prédictions d'une autre variable, à travers des chocs exogènes.



Graphique 14 : Décomposition de la variance des erreurs de prédiction

La plupart des séries économiques sont intégrées, ce qui a mené à un problème de régression fallacieuse[21] dans beaucoup d'analyses économiques. Le VECM a été développé, entre autres, pour éviter la différenciation nécessaire à l'utilisation des modèles ARMA. Ceux-ci occultent la dynamique de long-terme, en niveaux, au profit de celle à court-terme, en différences[22].

J'ai trouvé intéressant le cadre théorique du VECM, ses outils d'analyse, sa supposée robustesse à long-terme. Je l'ai donc sélectionné, avec l'ARIMAX, qui servira de témoin de par sa popularité.

A noter que l'on me demande des prédictions en nombre mais également en montant des entrées en impayés. J'ai décidé après quelques essais de faire l'hypothèse d'un montant moyen d'un impayé stable dans le futur, à partir de la valeur moyenne du dernier semestre connu. Cela améliore les prévisions en montant et permet de se concentrer sur la prévision en nombre.

Finalement, une technique prometteuse d'intelligence artificielle, les réseaux de neurones récurrents, se développent récemment dans le panorama des techniques de prévision. Mais cela n'était pas faisable directement sur le serveur de l'entreprise. J'ai cependant pu effectuer un test *ad-hoc* sur lequel je reviendrais en conclusion.

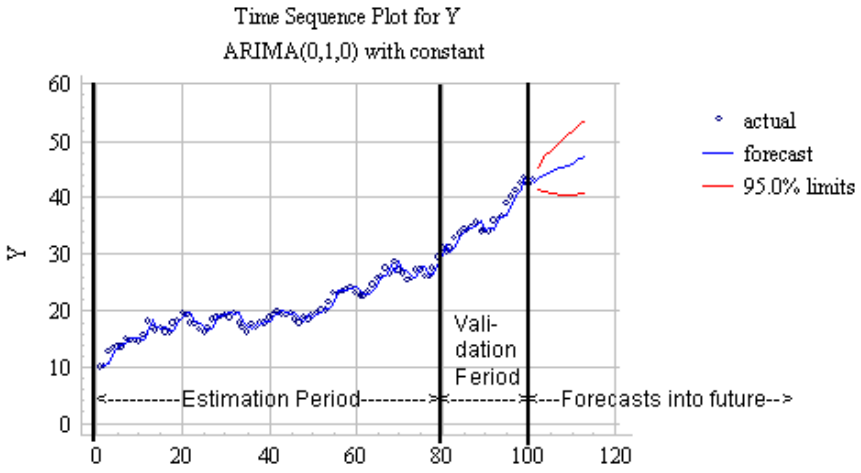
Les logiciels à ma disposition sont SAS Enterprise Guide 4.1 lié à un serveur d'entreprise, et Python 3 avec une distribution Anaconda locale. J'ai travaillé sous SAS pour estimer les VECM et ARIMAX et ainsi avoir accès au serveur. C'est sous SAS que j'ai effectué la validation et la prévision pour l'année budgétaire en cours. Concernant Python, sous lequel sont implémentées les méthodes les plus pointues de réseaux de neurones récurrents (avec les librairies TensorFlow et Keras), j'ai effectué le test *ad-hoc*.

Une migration des données d'un serveur à un autre a également été source de valeurs aberrantes, ce qui a amené la nécessité de faire un nettoyage de certains segments d'activité. Au mois de migration, une variation importante des valeurs était présente, hors de toute tendance, que j'ai décidé de mettre à zéro pour minimiser l'impact de la migration. Les modèles avaient en effet tendance à s'ajuster avec quelques mois de retard à cause de la migration non-corrigée.

Une étape importante de toute modélisation est la validation des modèles. Cela permet, dans le cadre d'une prévision, de comparer le pouvoir prédictif des modèles, et d'en effectuer la sélection, en amont de la prévision réelle. Ainsi, pour chaque segment, j'ai déterminé quel modèle retenir. Lors de la prévision effective, j'ai relancé l'estimation du modèle avec possible changement de variables, mais en gardant le même modèle.

Pour des données temporelles, la validation consiste donc à garder les données connues les plus récentes en dehors de la base d'apprentissage, dans un échantillon de validation. Cet échantillon de validation n'est pas utilisé lors de l'estimation. Cela permet d'évaluer les prévisions faites à ces dates par rapport aux données réelles avant la prévision effective et son évaluation *a posteriori*.

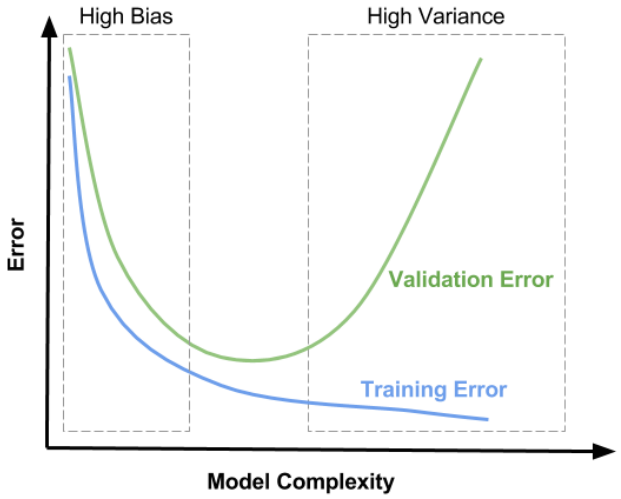
[21] http://cowles.yale.edu/sites/default/files/files/pub/d07/d0757.pdf
[22] Sargan, J. D. (1964). "Wages and Prices in the United Kingdom: A Study in Econometric Methodology", 16, 25–54. in *Econometric Analysis for National Economic Planning*



Graphique 15 : Validation classique en séries temporelles[23]

Une erreur de validation peut être calculée, qui évolue différemment de l'erreur sur l'échantillon d'estimation, en fonction de la complexité du modèle. Tandis que l'erreur d'estimation baisse continuellement selon la complexité du modèle, l'erreur de prédiction remonte à un certain moment.

C'est un problème d'*overfitting* (non parcimonie) : le modèle s'améliore seulement sur les données qu'il connait.



Graphique 16 : Erreurs d'estimation et de validation[24]

[23] https://people.duke.edu/~rnau/three.htm
[24] https://towardsdatascience.com/cross-validation-a-beginners-guide-5b8ca04962cd

Dans une optique prédictive, la métrique d'évaluation des modèles évalués se doit d'être une erreur de prédiction. Il en existe différentes[25] :

$$MAE(Mean\ Absolute\ Error) = \frac{1}{h}\sum_{j=1}^{h}|y_{t+j} - \hat{y}_{t+j}|$$

$$MAPE(Mean\ Absolute\ Percentage\ Error) = \frac{1}{h}\sum_{j=1}^{h}\frac{|y_{t+j} - \hat{y}_{t+j}|}{y_{t+j}}$$

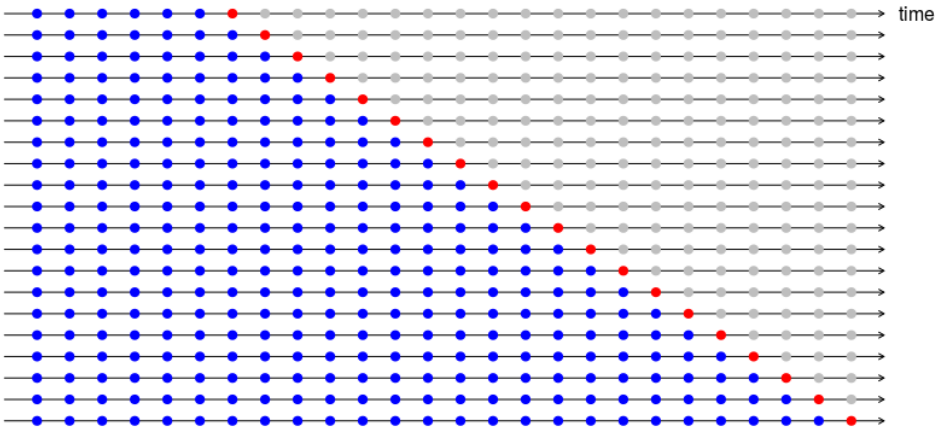
$$sMAE(Scaled\ Mean\ Absolute\ Error) = \frac{MAE}{\bar{y}}$$

$$MASE(Mean\ Absolute\ Scaled\ Error) = \frac{MAE}{\frac{1}{t-1}\sum_{j-2}^{t}|y_j - y_{j-1}|}$$

$$RMSE(Root\ Mean\ Squared\ Error) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

MAPE, sMAE, peuvent s'interpréter en pourcentage d'erreur, donc en absolu. Les autres erreurs sont des erreurs relatives, nécessitant donc une comparaison. Finalement, l'écart absolu au réel semestriel a été retenu, que ce soit en nombre ou en montant. Il faut ainsi agréger les données mensuelles par semestre avant de calculer l'erreur absolue. L'objectif principal est alors de limiter cette erreur à 10% à des horizons d'un semestre, deux semestres et trois semestres.

Finalement, une méthode récente issue du *machine learning*, qui aurait pu être mise en place, s'appelle la *cross validation.* Cela consiste à créer plusieurs échantillons de validation par coulissement de la série dans le temps.



Graphique 17 : schéma de cross validation (bleu : échantillon d'estimation, rouge : échantillon de validation) [26]

[25] http://forecasting.svetunkov.ru/en/2017/07/29/naughty-apes-and-the-quest-for-the-holy-grail/
[26] https://robjhyndman.com/hyndsight/tscv/

Les erreurs de validation sont alors moyennées pour donner l'erreur de *cross validation*. Cela permet d'avoir une erreur de validation plus robuste et donc de trouver un modèle mieux paramétré.

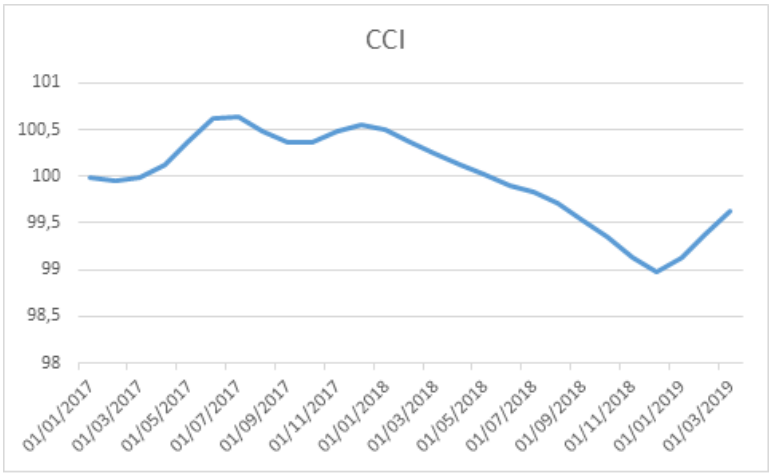
N'ayant pas beaucoup de données sur lesquelles travailler, il m'a fallu effectuer une validation classique :

Deuxième partie

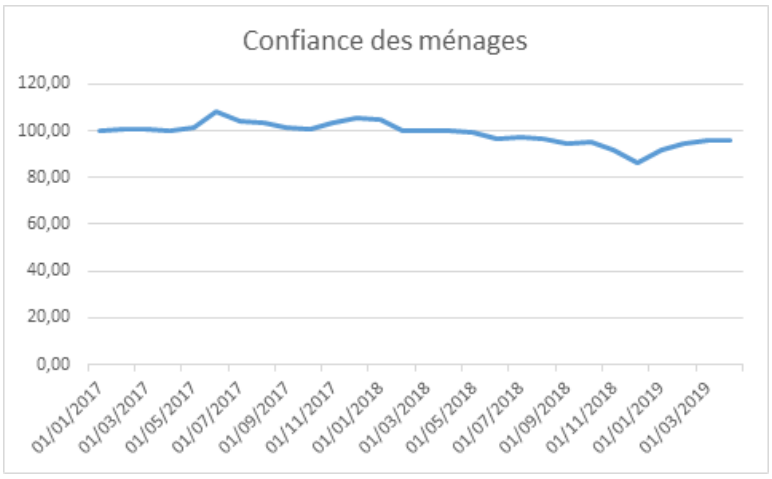
Présentation des variables externes

Voici l'évolution de janvier 2017 à la date la plus récente des indicateurs macroéconomiques, qui sont, je le rappelle, publics.

Le *Confidence Consumer Index* de l'OCDE et la confiance des ménages de l'INSEE évoluent de manière similaire, cependant le CCI a des tendances plus marquées.

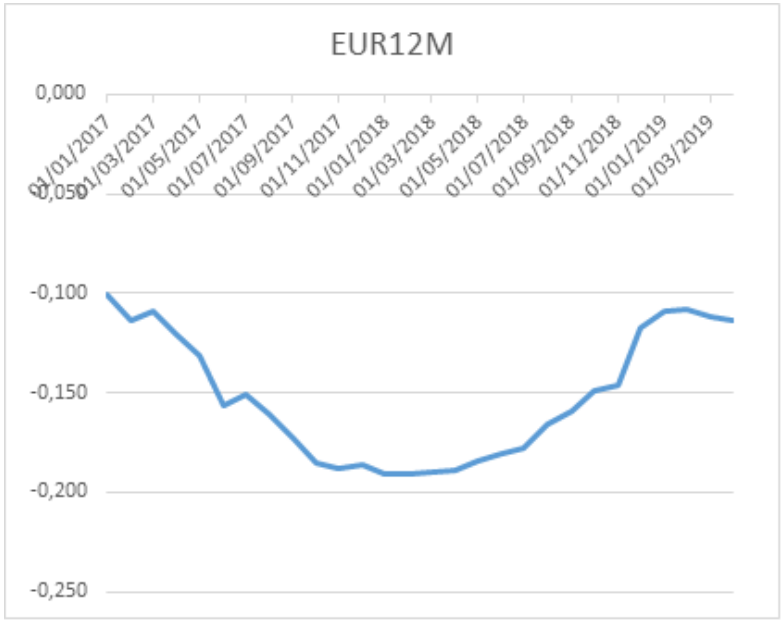


Graphique 18 : Historique du *Consumer Confidence Index* selon l'OCDE

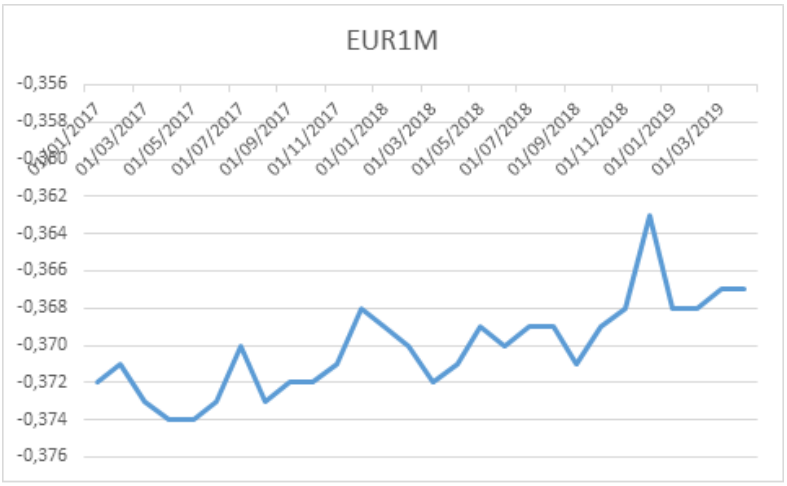


Graphique 19 : Historique de la Confiance des ménages français selon l'INSEE

L'Euribor 12 mois et l'Euribor 1 mois ont une évolution différente. Alors que l'EUR12M semble converger vers une valeur historique, l'EUR1M est en augmentation constante. Ce dernier semble aussi plus volatil.

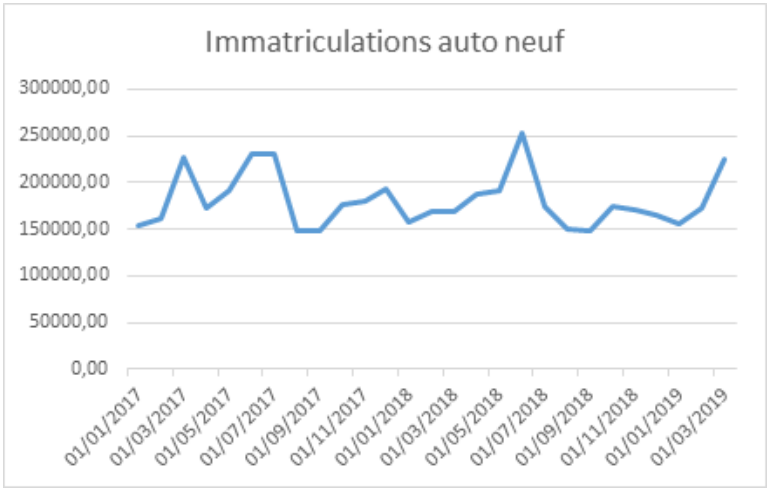


Graphique 20 : Historique de l'Euribor 12 mois selon la Banque de France

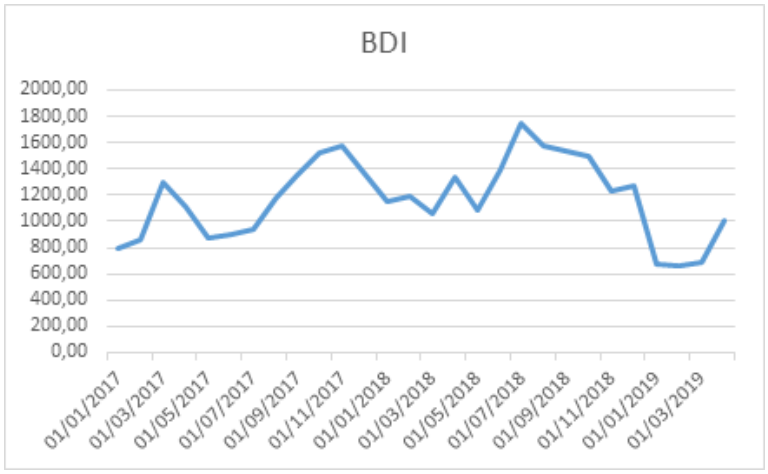


Graphique 21: Historique de l'Euribor 1 mois selon la Banque de France

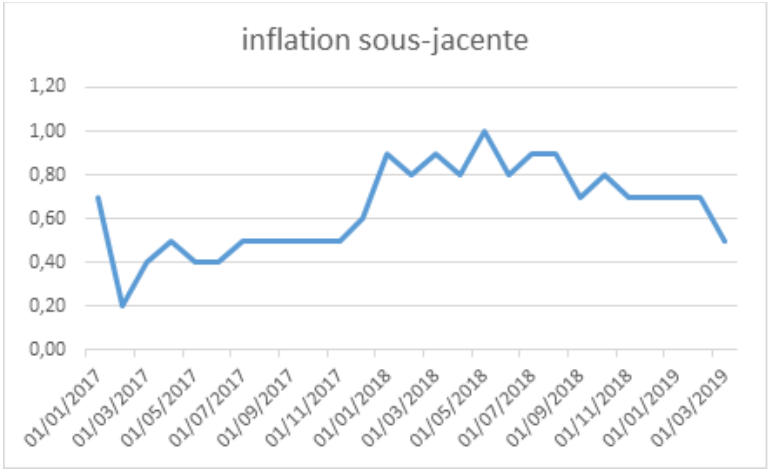
Les immatriculations neuves françaises et le *Baltic Dry Index* ont un pic à la mi-2018 avant de tomber lors de l'hiver 2018. L'inflation est encore stable malgré la promesse de remontées prochaines, que l'on observe dans les immatriculations et le BDI.



Graphique 22: Historique des immatriculations neuves françaises selon le CCFA

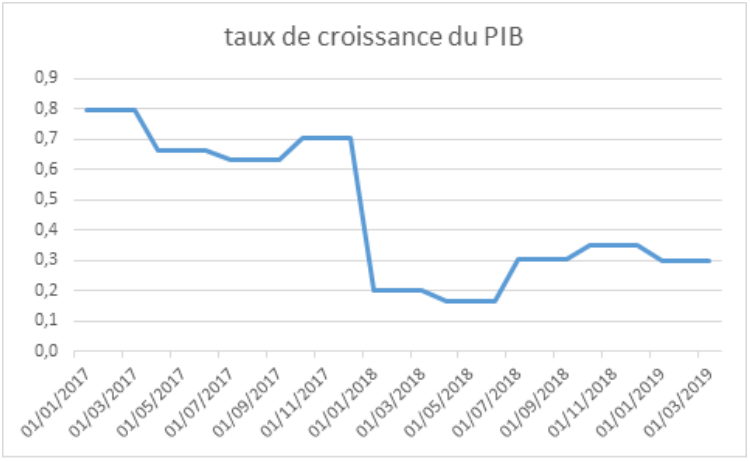


Graphique 23: Historique du *Baltic Dry Index* selon Investing.com

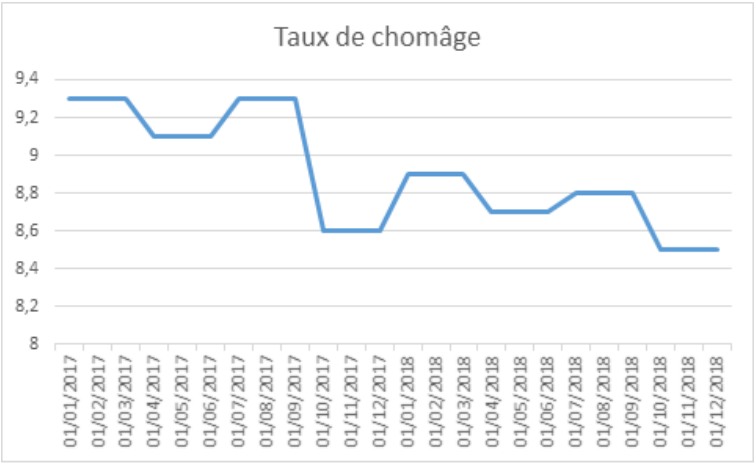


Graphique 24: Historique de l'inflation sous-jacente selon l'INSEE

Le PIB et le chômage connaissent un décalage vers la fin de l'année 2017. En dehors de cette période, les niveaux restent stables.

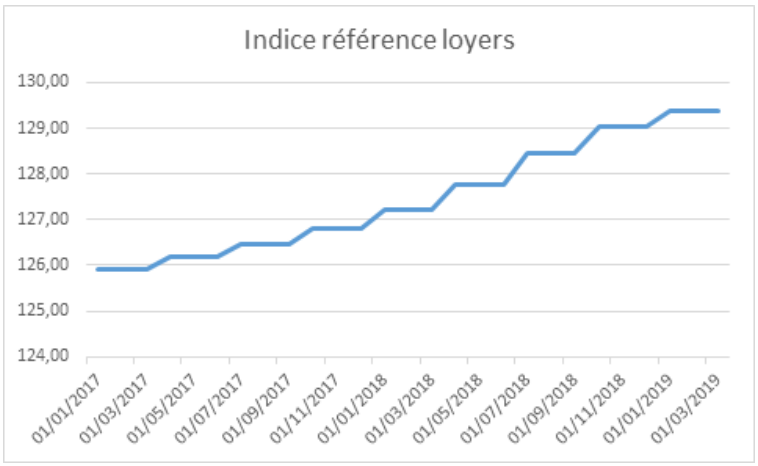


Graphique 25: Historique du taux de croissance du PIB français selon l'INSEE

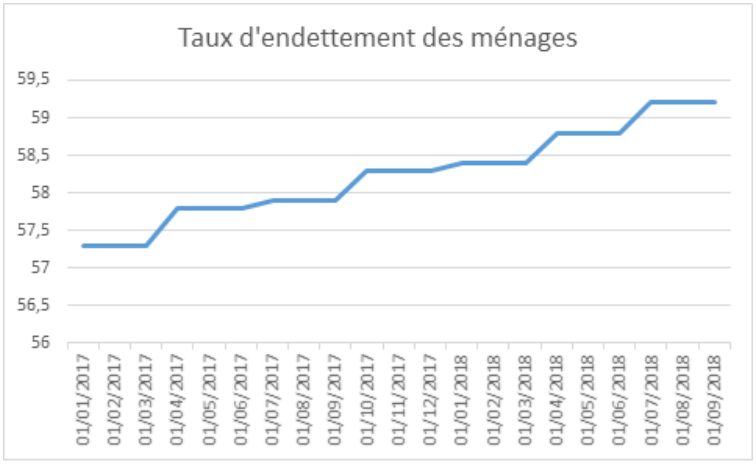


Graphique 26: Historique du taux de chômage français français selon l'INSEE

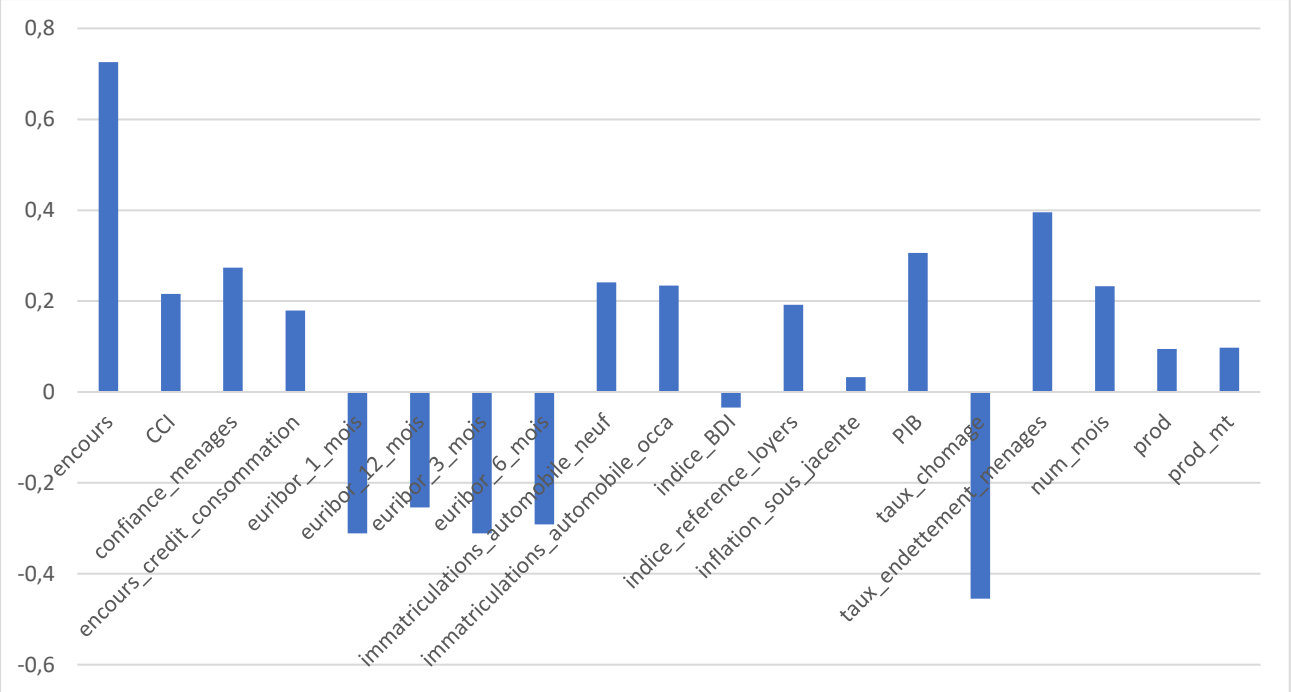
L'indice de référence des loyers et le taux d'endettement des ménages croissent, eux, continuellement.



Graphique 27: Historique de l'indice de référence des loyers selon l'INSEE



Graphique 28: Historique du taux d'endettement des ménages français selon la Banque de France



Graphique 29: Corrélations croisées avec impayés contemporaines

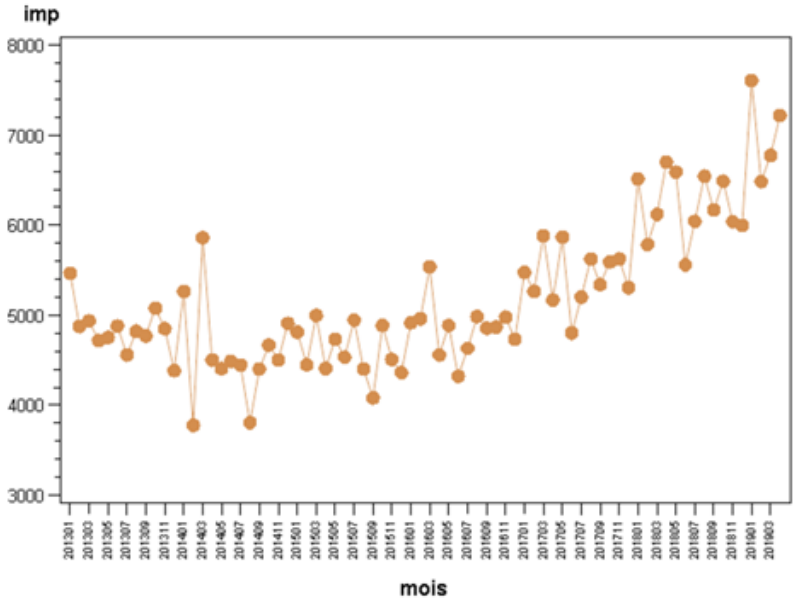
Finalement, il semble donc que certains indicateurs caractérisent les mêmes phénomènes sous-jacents. Ils seront certainement redondant dans la modélisation.

Pratique de la modélisation

Pour l'estimation des modèles, j'ai d'abord commencé par l'ARIMAX avec une régression pas à pas. J'intègre toutes les variables (après *prewhitening*), que j'élimine une par une suivant leur significativité. J'arrête la procédure lorsque toutes les variables restantes sont significatives à un seuil de 5% de risque. Je teste après cela l'ajout de retards sur les variables, en regardant l'AIC (*Akaike Information Criterion*). Ensuite, je vérifie la non-significativité des autocorrélations des résidus, que je corrige éventuellement en ajoutant des termes ARMA.

J'ai ensuite utilisé les variables choisies dans le VECM. Comme il faut y trouver une relation de cointégration intéressante, j'ai pu changer les variables qui ne donnaient pas une relation intéressante, pour d'autres. Une fois qu'une relation de cointégration satisfaisante est trouvée dans le VECM grâce au test de la trace de Johansen, l'ordre autorégressif est choisi par *grid search* sur l'AIC. Celui-ci doit rendre un résultat parcimonieux, donc généralisable *out of sample.*

Voici un exemple d'estimation et de prévision sur un segment quelconque, avec ARIMAX et VECM. La série des nombres d'impayés figure ci-dessous :



Graphique 30: Historique du nombre d'entrées en impayés sur un segment quelconque

La sélection pas à pas donne le modèle suivant. A noter que deux variables dépassent le seuil de 5% de significativité mais permettent une baisse de l'AIC, que je privilégie.

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
AR1,1	-0.84512	0.12765	-6.62	<.0001	1	imp	0
AR1,2	-0.41427	0.12905	-3.21	0.0023	2	imp	0
NUM1	2.86780	17.27161	0.17	0.8688	0	CCI	12
NUM2	0.0008654	0.00009184	9.42	<.0001	0	encours	0
NUM1,1	0.0008580	0.00009531	9.00	<.0001	1	encours	0
NUM3	-5.60581	35.67829	-0.16	0.8758	0	taux_endettement_menages	0

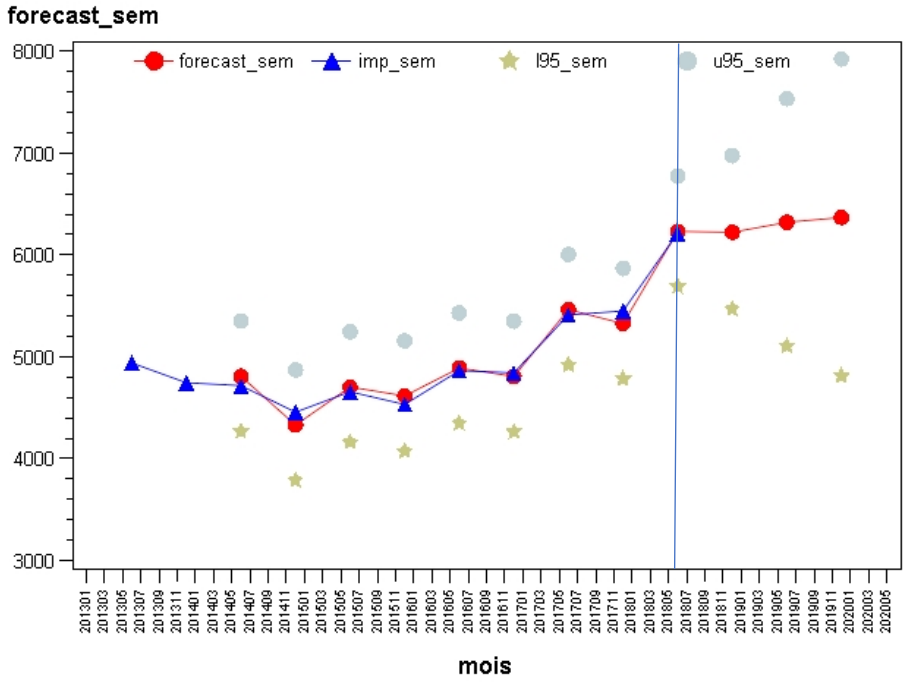
Table 2 : coefficients d'un ARIMAX estimé sur un segment

Les résidus n'ont pas d'autocorrélation significative 24 mois en arrière (les autocorrélations croisées résidus/séries externes sont aussi non-significatives).

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	2.78	4	0.5961	-0.083	-0.128	-0.131	0.011	-0.067	-0.005
12	10.87	10	0.3677	0.151	-0.016	-0.016	-0.138	-0.223	0.142
18	15.88	16	0.4613	-0.036	-0.015	-0.002	0.119	0.023	-0.207
24	19.79	22	0.5964	0.065	0.004	-0.027	0.043	0.072	0.165

Table 3 : autocorrélations d'un ARIMAX estimé sur un segment

Les prédictions à 6,12 et 18 mois, donc agrégées semestriellement, figurent ci-dessous. L'intervalle de confiance croit linéairement dans le temps, la prévision est plutôt incertaine.



Graphique 31: Prévision des nombres d'entrées en impayé, effectuée par ARIMAX à 6, 12 et 18 mois, avec intervalles de confiance à 95%

L'estimation d'un VECM nécessite des séries non-stationnaires pour appliquer la cointégration. Les tests de Dickey-Fuller suivants le confirment bien :

Dickey-Fuller Unit Root Tests					
Variable	Type	Rho	Pr < Rho	Tau	Pr < Tau
imp	Zero Mean	0.32	0.7564	0.64	0.8528
	Single Mean	-3.11	0.6362	-0.81	0.8102
	Trend	-19.57	0.0562	-3.06	0.1237
encours	Zero Mean	0.01	0.6816	0.01	0.6841
	Single Mean	-5.03	0.4207	-1.39	0.5848
	Trend	-12.43	0.2617	-2.89	0.1707
euribor_1_mois	Zero Mean	0.28	0.7464	0.27	0.7615
	Single Mean	-1.03	0.8805	-0.86	0.7965
	Trend	-3.55	0.9072	-1.13	0.9164

Table 4 : test de Dickey-Fuller pour VECM estimé sur un segment

Le test de la trace de Johansen suivant donne un rang de cointégration de 1.

Cointegration Rank Test Using Trace						
H0: Rank=r	H1: Rank>r	Eigenvalue	Trace	5% Critical Value	Drift in ECM	Drift in Process
0	0	0.2461	31.6907	29.38	Constant	Linear
1	1	0.1178	11.0669	15.34		
2	2	0.0259	1.9171	3.84		

Table 5 : test de la trace de Johansen pour VECM estimé sur un segment

Finalement, l'AIC conseille un ordre autorégressif de 3.

Minimum Information Criterion			
Lag	MA 0	MA 1	MA 2
AR 0	35.501741	36.058917	36.217503
AR 1	31.211045	31.233932	31.477357
AR 2	30.932301	31.282297	31.721277
AR 3	30.785954	31.203037	31.499851
AR 4	30.959536	31.495444	31.917583
AR 5	31.496243	32.033514	32.444921
AR 6	32.090704	32.942055	33.54386

Table 6 : grille de critère d'information par ordre autorégressif pour VECM estimé sur un segment

On estime ainsi le VECM avec les paramètres ci-dessous. L'EUR1M influence les nombres d'impayés (imp) mais pas les montants (encours), qui sont indépendants. Cela dénote d'une mauvaise spécification des montants. J'aurais pu les enlever et les remplacer par une variable macroéconomique. L'hypothèse de montant moyen stable dans les 18 mois de prédiction permet de s'en passer.

Schematic Representation of Parameter Estimates				
Variable/Lag	C	AR1	AR2	AR3
imp	+	***	-..	-..
encours	+	***	.-.	...
euribor_1_mois	.	***	..-	+.-
+ is > 2*std error, - is < -2*std error, . is between, * is N/A				

Table 7 : représentation schématique des coefficients estimé d'un VECM

Voici la relation de long-terme (de cointégration) entre variables en niveaux et les ajustements de court-terme en différences.

Long-Run Parameter Beta Estimates When RANK=1	
Variable	1
imp	-0.00226
encours	0.00000
euribor_1_mois	2.08306

Adjustment Coefficient Alpha Estimates When RANK=1	
Variable	1
imp	-24.68371
encours	-91286.93111
euribor_1_mois	0.01163

Table 8 : paramètres de long et court termes estimés d'un VECM

Des indicatrices saisonnières peuvent être mises en place. Les données d'impayés sont bien saisonnières mais pas l'Euribor.

Schematic Representation of Seasonal Dummy Estimates											
Variable/Lag	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11
imp	-	.	-	-	-	-	-	-	.	-	-
encours	-	.	-	-	-	-	-	-	-	-	-
euribor_1_mois	.	.	.	.	.	.	.	.	.	.	.
+ is > 2*std error, - is < -2*std error, . is between, * is N/A											

Table 9 : indicatrices saisonnières estimées d'un VECM

Il reste des corrélations croisées significatives, toutefois je n'ai pas changé l'ordre autorégressif choisi par AIC.

Schematic Representation of Cross Correlations of Residuals

Variable/Lag	0	1	2	3	4	5	6	7	8	9	10	11	12
imp	++.	...	...	...	...	..+	...	...	...	...	--.	...	...
encours	++.	...	...	...	...	...	...	...	...	...	-..	...	...
euribor_1_mois	..+	...	...	...	...	...	...	...	...	...	...	...	...

+ is > 2*std error, - is < -2*std error, . is between

Table 10 : corrélations croisées estimées d'un VECM

Analyse des résultats

Les outils d'analyse vus auparavant sont ci-dessous. La proportion de variance d'erreur de prédiction sur le nombre d'impayés portée par l'Euribor se situe autour de 1%. Les fonctions de réponse aux impulsions sont présentées seulement de manière orthogonalisée.

Proportions of Prediction Error Covariances

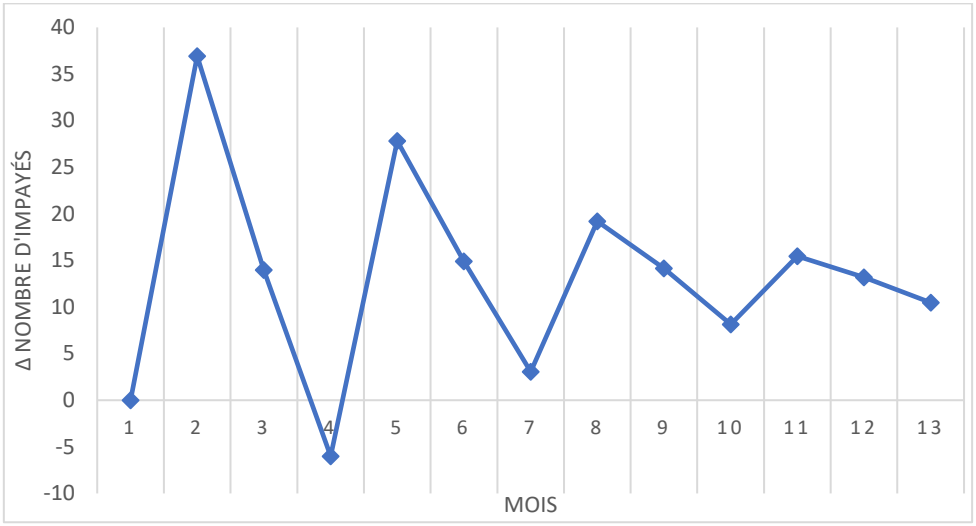
Lead	Variable	imp	encours	euribor_1_mois
1	imp	1.00000	0.00000	0.00000
	encours	0.68831	0.31169	0.00000
	euribor_1_mois	0.00378	0.00000	0.99622
2	imp	0.98678	0.00123	0.01199
	encours	0.70422	0.28301	0.01278
	euribor_1_mois	0.01516	0.00472	0.98012
3	imp	0.98407	0.00318	0.01276
	encours	0.67437	0.31210	0.01353
	euribor_1_mois	0.03344	0.00578	0.96079
4	imp	0.98826	0.00231	0.00943
	encours	0.73094	0.25826	0.01079
	euribor_1_mois	0.02662	0.02084	0.95253

Orthogonalized Impulse Response

Lag	Variable	imp	encours	euribor_1_mois
0	imp	322.87638	0.00000	0.00000
	STD	26.72144	0.00000	0.00000
	encours	244916.55440	164809.79564	0.00000
	STD	54.81714	43761.27224	0.00000
	euribor_1_mois	-0.00199	0.00003	0.03227
	STD	37.82725	34663.63698	0.00268
1	imp	88.83294	11.80827	36.92015
	STD	32.16990	37.76295	36.21266
	encours	89256.72713	12056.17690	35115.24064
	STD	29515.24514	35116.67474	33407.78192
	euribor_1_mois	0.00435	0.00267	0.02088
	STD	0.00317	0.00359	0.00391

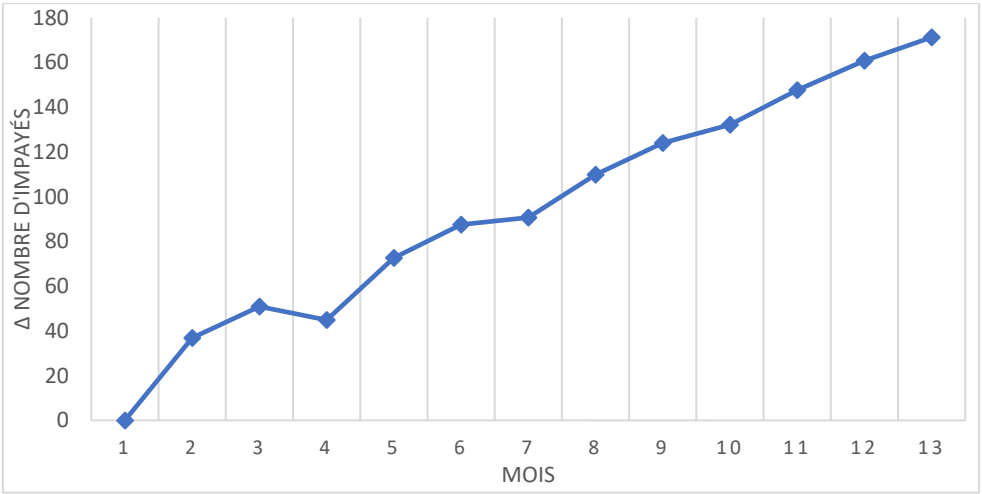
Tables 11-12 : proportions de variance d'erreur de prédiction (à gauche), fonctions de réponse impulsionnelle orthogonalisées (à droite) d'un VECM

La réponse des impayés à un choc de l'Euribor 1 mois orthogonal est la suivante. Après l'initialisation du choc en t1, la conséquence du choc est suivie par mois. Celle-ci est positive après un an, le choc n'est pas résorbé, cela tendant tout de même vers 0.



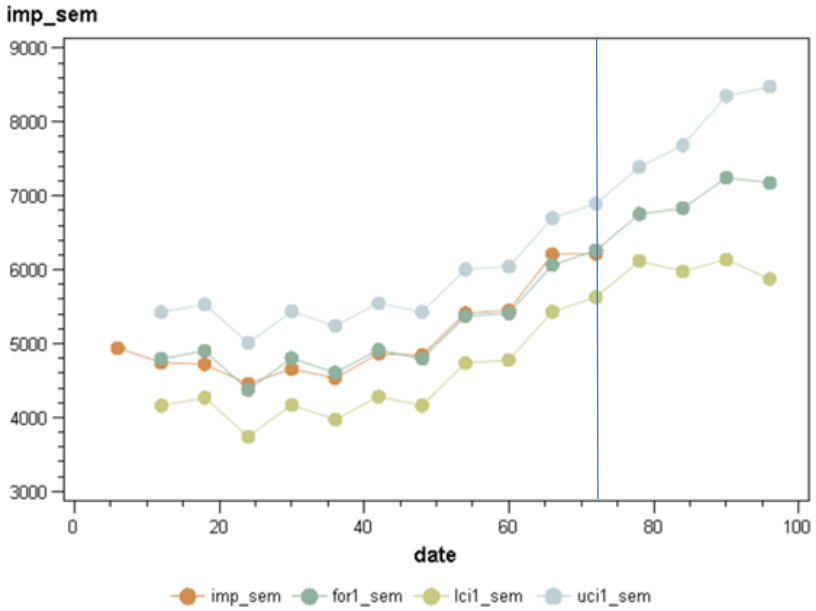
Graphique 32 : fonction de réponse impulsionnelle des nombres d'impayés à l'Euribor 1 mois

En cumulant les impacts mensuels, la réponse a été positive.



Graphique 33 : fonction de réponse impulsionnelle cumulée des nombres d'impayés à l'Euribor 1 mois

Voici les prévisions effectuées ci-dessous. La tendance est plus haussière que dans l'ARIMAX. L'intervalle de confiance est aussi plus réduit.



Graphique 34 : Prévisions des nombres d'entrées en impayé semestriels par VECM, avec intervalles de confiance à 95%

(Données arrêtés en avril 2019 donc 4 semestres sont prédits, en fait 20 mois)

Finalement, le récapitulatif des résultats de validation figure en dessous.

		REEL			
	semestre	nb_reel	mt_reel	mt_moy_reel	mt_moy_hyp
APPRENTISSAGE	1S2017	32463	31103627	958	958
	2S2017	32688	30214945	924	958
	1S2018	37270	34570935	928	958
TEST	2S2018	37283	35027151	939	958

ESTIMATION ARIMAX				
nb_prevu	mt_prevu	ecart_nb	%nb	%mt
32794	31420809	794	1,02%	1,02%
31778	30447425	2880	-2,78%	0,77%
32727	31357014	9187	-12,19%	-9,30%
32090	30745955	13086	-13,93%	-12,22%

ESTIMATION VECM								
nb_prevu	mt_prevu	ecart_nb	ecart_mt	%nb	%mt	mt_moy	mt_prev2	%mt2
32013	29933472	2094	2129329	-1,39%	-3,76%	935,0485	30672229	-1,39%
33525	31540531	1636	1302296	2,56%	4,39%	940,8173	32120779	6,31%
34243	31680492	5488	3814992	-8,12%	-8,36%	925,169	32809017	-5,10%
34837	31639368	4980	4013760	-6,56%	-9,67%	908,2007	33378618	-4,71%

Table 13 : Récapitulatif des données de la validation effectuée

4 semestres sont représentés : le dernier de l'échantillon d'apprentissage et les 3 semestres de validation. En vert figurent les données réelles, en orange les prévisions (ARIMAX puis VECM) et en jaune les erreurs (écarts au réel, nombre et montant, en pourcentage).

Seul le VECM reste en dessous de 10% d'écart au réel en nombre. En montant, il faut l'hypothèse de montant moyen (*%mt2*) pour préserver cet objectif.

Sur tous les segments, le seuil d'erreur de 10% n'a pas été dépassé.

Le modèle le plus fréquemment choisi est le VECM. Avec l'hypothèse de montant moyen stable dans la période de prévision, sa performance est souvent améliorée. L'ARIMAX se fait distancer à long-terme tandis qu'il reste assez proche à court-terme.

Les différents Euribor et indices de confiance sont les deux données qui sont les plus utilisées.

Lors de la phase de prédiction effective, en avril 2019, les résultats que j'ai obtenus ont été similaires à ceux de ma tutrice. Sa méthode est toutefois plus laborieuse, car elle nécessite d'être ajustée. Alors que cela prend normalement plusieurs jours, les prévisions peuvent être faites en quelques heures avec ma méthode. Les prévisions ont été bien reçues par ma tutrice ainsi que par les différents responsables. Le pouvoir explicatif du VECM a été un point qui intéresse toute la hiérarchie, notamment l'idée d'effectuer un stress-test.

La procédure SAS qui permet l'estimation du VECM est aussi plus ergonomique que celle de l'ARIMAX, qui doit passer par une phase laborieuse de *prewhitening* des séries univariées, soit estimer un processus ARMA pour chaque série. Le VECM a aussi l'avantage d'être multivarié, donc de pouvoir effectuer des prévisions sur n'importe quelle variable temporelle le composant.

Au final, je préconiserais donc, à l'avenir, pour mon successeur, de continuer à étudier le VECM. Lors des prochaines phases de prévisions, l'ARIMAX peut être utilisé pour constater l'écart qu'ont les deux méthodes. En général, le sens de variation est au moins similaire, dans le cas contraire, il y aurait matière à revoir le processus.

Le processus peut aussi être automatisé, notamment pour sortir des graphiques, qui manquent cruellement dans la version de SAS de l'entreprise. ODS Graphics n'est en effet pas implémenté. Cela passera sûrement par des macros SAS et des applications stockées.

Il pourrait y avoir lieu, comme je l'ai évoqué auparavant, de faire plus attention à la préparation des données. La dessaisonalisation, la remise à la bonne fréquence, sont autant de sujets potentiels. La recherche de variables étant cointégrées avec les impayés est aussi toujours intéressante.

Conclusion

Récapitulatif

Cette étude avait comme enjeu de produire des prévisions des entrées en impayés, de manière robuste dans le long-terme, fondée statistiquement et incorporant des données externes. Les données sont mensuelles, les prédictions sont agrégées en semestres ; les trois semestres futurs ont ainsi été prévus.

Des données issues de l'Open Data ont été utilisées, comme l'Euribor à un an ou le *Consumer Confidence Index*.

A l'aide de modèles statistiques de séries temporelles, comme le VECM, j'ai pu mettre en évidence une dynamique reliant ces données aux impayés.

A long terme, une hypothèse de montant moyen permet de mieux modéliser les montants globaux. On se concentre ainsi sur les nombres d'impayés lors d'une première étape.

La méthode appliquée permet d'analyser la dynamique et l'interaction qu'ont les variables entre elles, à travers la décomposition de la variance de l'erreur de prédiction et les fonctions de réponse aux impulsions.

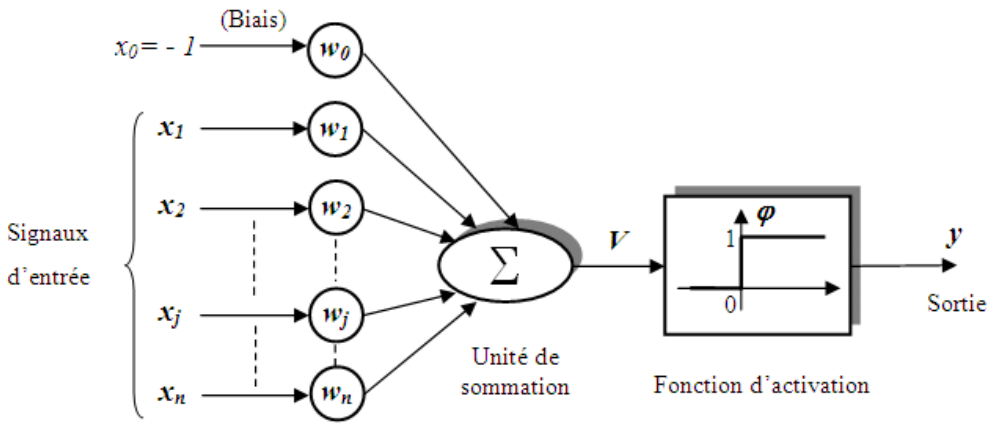
La validation a permis de dégager les modèles les plus performants par segment. Lors de la prévision pour le budget 2020, mes prévisions ont été mises en place rapidement et ont été en accord avec celles de ma tutrice. Le directeur a accueilli ces prévisions de bonne manière. Cela a ainsi été une réussite. Une présentation au comité exécutif de l'entreprise est même prévue.

Il reste maintenant à automatiser la procédure et continuer à étudier les outils d'analyse explicative.

Les réseaux de neurones récurrents

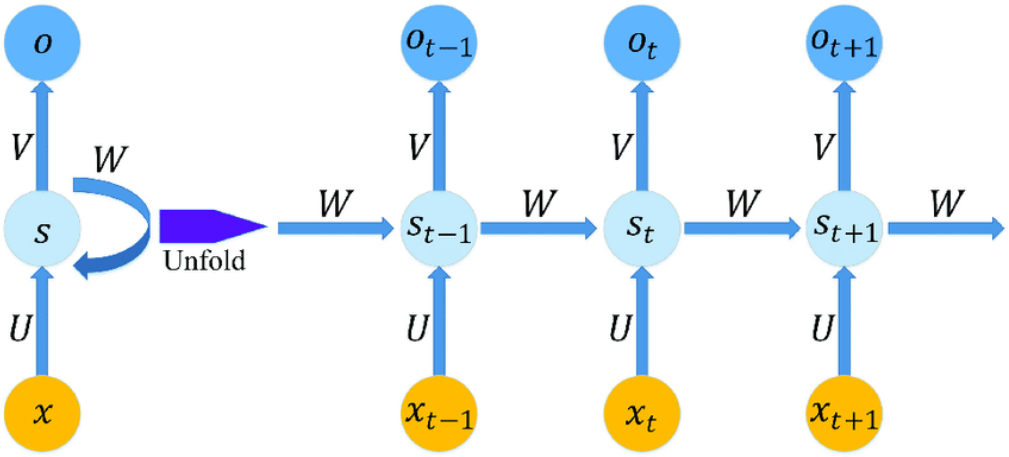
Finalement, avec l'avènement de l'Intelligence Artificielle et du Machine Learning, des méthodes qui ne sont pas implémentées dans la version SAS d'entreprise apparaissent. Dans la littérature scientifique, les réseaux de neurones sont très appréciés. Ils ont la propriété d'approximateurs universels de fonctions[27]. Des relations non-linéaires, que ne détectent pas les méthodes traditionnelles, pourraient améliorer la précision de prévision. Cela est dû aux fonctions d'activation des neurones.

[27] Cybenko, G.V. (2006). "Approximation by Superpositions of a Sigmoidal function". In van Schuppen, Jan H. (ed.). Mathematics of Control, Signals, and Systems. Springer International. pp. 303–314.



Graphique 35 : prototype d'un neurone artificiel[28]

Pour la prévision de séries temporelles, une architecture spécifique existe, les réseaux de neurones récurrents. Ils permettent de modéliser une information séquentielle :



Graphique 36 : un réseau de neurones récurrent, dont vue déployée à droite[29]

U, V et W sont les poids de la couche cachée, de la couche d'output et de l'état interne s_t. Les vecteurs x et o sont les input et output au temps t.

J'ai effectué un test en Python, avec Keras, d'un réseau de neurones artificiels récurrent. Cela demande d'ajuster un certain nombre de paramètres :

- Type de modélisation multivariée (encodeur-décodeur, output vectoriel)
- Type de neurone (simple, *LSTM*)
- Type de fonction d'activation (ReLU, TanH, *etc.*)

[28] Benharir, N & Zerikat, M & Chekroun, Soufyane & Mechernene, Abdelkader. (2014). Approche Adaptative d'une Commande Neuronale sans capteur d'un Moteur Asynchrone associée à un Observateur par Mode Glissant.

[29] Bao, Wei & Yue, Jun & Rao, Yulei. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE. 12. 10.1371/journal.pone.0180944.

- Type d'erreur (erreur moyenne quadratique, *etc*.)
- Nombre de neurones
- Nombre de couches
- Taille de *batch*
- Nombre d'*epochs*
- Technique de régularisation (*dropout*)
- Technique de *preprocessing* (dessaisonalisation, normalisation)

Le résultat n'a pas été totalement concluant, malgré des promesses intéressantes. Cette méthode est, de toute façon, une boîte noire, donc sans explications à donner sur ses résultats. De plus, un serveur Python n'est pas encore en marche dans l'entreprise. Je n'ai donc pas poursuivi cela outre-mesure.

Ce rapport touche à sa fin. Je tiens à dire que cette année m'aura apporté beaucoup de connaissances en statistiques. J'ai pu appliquer mes connaissances sur un projet intéressant et lié aux grandes problématiques de la direction. J'ai pu grandir au sein du monde de l'entreprise et me préparer à une nouvelle année d'études, à l'ENSAI.

Ce sera l'occasion pour moi de mieux encore connaître la science des données.

Bibliographie

https://freakonometrics.hypotheses.org/tag/chain-ladder

https://business.lesechos.fr/entrepreneurs/marketing-vente/0301760163621-organiser-un-salon-un-booster-commercial-pour-quantmetry-322589.php

https://www.banque-france.fr/statistiques/credit/credit/credits-la-consommation

https://www.lesechos.fr/finance-marches/banque-assurances/lannee-2019-a-mal-commence-pour-le-marche-du-credit-a-la-consommation-1001973

https://www.lesechos.fr/finance-marches/banque-assurances/banques-les-nouvelles-techniques-des-cyber-braqueurs-959488

https://www.thalesgroup.com/fr/marches-specifiques/systemes-dinformation-critiques-et-cybersecurite/news/thales-et-sekoia

https://www.oecd.org/sdd/compositeleadingindicatorsclifrequentlyaskedquestionsfaqs.htm#1

https://www.insee.fr/fr/metadonnees/definition/c1599

https://www.math.univ-toulouse.fr/~lagnoux/Poly_SC.pdf

https://stats.idre.ucla.edu/sas/seminars/sas-survival/

https://www.ajol.info/index.php/orion/article/download/111549/101328

http://bresson.u-paris2.fr/doku.php?id=start

https://robjhyndman.com/hyndsight/arimax/

http://mayoral.iae-csic.org/timeseries2019/handout2_arma.pdf

http://www-personal.umich.edu/~franzese/DeBoefKeele.2008.TakingTimeSeriously.pdf

http://statmath.wu.ac.at/~hauser/LVs/FinEtricsQF/FEtrics_Chp4.pdf

http://cowles.yale.edu/sites/default/files/files/pub/d07/d0757.pdf

https://people.duke.edu/~rnau/three.htm

https://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique

https://towardsdatascience.com/cross-validation-a-beginners-guide-5b8ca04962cd

http://forecasting.svetunkov.ru/en/2017/07/29/naughty-apes-and-the-quest-for-the-holy-grail/

https://robjhyndman.com/hyndsight/tscv/

https://machinelearningmastery.com/

https://towardsdatascience.com/

https://stats.stackexchange.com/

https://support.sas.com

Bao, Wei & Yue, Jun & Rao, Yulei. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLoS ONE. 12. 10.1371/journal.pone.0180944.

Benharir, N & Zerikat, M & Chekroun, Soufyane & Mechernene, Abdelkader. (2014). Approche Adaptative d'une Commande Neuronale sans capteur d'un Moteur Asynchrone associée à un Observateur par Mode Glissant.

Cybenko, G.V. (2006). "Approximation by Superpositions of a Sigmoidal function". In van Schuppen, Jan H. (ed.). Mathematics of Control, Signals, and Systems. Springer International. pp. 303–314.

Gagniuc, Paul A. (2017). *Markov Chains: From Theory to Implementation and Experimentation*. USA, NJ: John Wiley & Sons. pp. 1–256. ISBN 978-1-119-38755-8.

George Edward Pelham Box and Gwilym Jenkins. 1990. Time Series Analysis, Forecasting and Control. Holden-Day, Inc., San Francisco, CA, USA.

Robert F. Engle and C. W. J. Granger, Co-Integration and Error Correction: Representation, Estimation, and Testing, Econometrica, Vol. 55, No. 2 (Mar., 1987), pp. 251-276

Sargan, J. D. (1964). "Wages and Prices in the United Kingdom: A Study in Econometric Methodology", 16, 25–54. in *Econometric Analysis for National Economic Planning*

Table des matières

La prévision de données temporelles à des horizons de 6 mois, 12 mois et 18 mois est toujours un sujet de discussion brûlant. A des méthodes statistiques traditionnelles s'ajoutent des méthodes plus récentes d'intelligence artificielle. Le sujet de ce rapport sont les entrées en impayé des crédits à la consommation de Crédit Agricole Consumer Finance (CACF). Leur prévision doit permettre d'établir le budget de l'année en cours pour la *Business Unit* France.

Dans ce rapport, je décris toutes les étapes d'une étude statistique, avec pour aboutissement la prévision effective des entrées en impayé, pour les 18 prochains mois.

Mots-clefs : prévision, VECM, ARIMA, RNN, impayés