Institut Louis Bachelier
DATALAB
Palais Brongniart, 28 place de la Bourse, 75002 Paris, France

Université de Technologie de Compiègne, Sorbonne Universités
Département de Génie Informatique
Rue du docteur Schweitzer, 60203 Compiègne, France

Rapport de stage TN10

Projets de recherche appliquée en data science

Etudiant stagiaire:
Anthony Galtier

Maître de stage:
Louis Boulanger

Suiveur de stage UTC:
Franck Davoine

Lundi 10 février 2020

Remerciements

Je tiens à remercier Louis Boulanger, directeur du Datalab à l'Institut Louis Bachelier et tuteur de mon stage ici pour m'avoir offert l'opportunité de rejoindre cette équipe, pour sa bienveillance au cours de ce stage et pour ses conseils précieux. Je remercie également Driss Lamrani pour son enthousiasme, ses conseils précieux et son implication tout au long de ces six mois. Je remercie aussi Krim Ziane pour son bon vouloir, ses qualités pédagogiques et son implication dans le travail réalisé. Je remercie tous les membres du Datalab pour leur bonne collaboration et le bon esprit d'équipe. Je remercie enfin Franck Davoine, mon tuteur de stage à l'UTC pour sa visite et sa disponibilité au cours de ce stage.

Sommaire

1 Confidentialité de certaines données et conformité au RGPD

Au cours de ce stage, j'ai été amené à travailler sur plusieurs ensembles de données de nature différente et dont certaines sont confidentielles. Certaines données exploitées dans le cadre du projet commissionné par le groupe Generali, en particulier l'extrait du portefeuille MRC mis à notre disposition, sont soumises à un accord de non divulgation. Il en est de même pour les relevés GPS fournis par Asment Temara ainsi que pour l'extrait de l'IDEX fourni par Everdiams. Ces données et toute analyse de celles-ci ne seront donc pas détaillées dans ce rapport. Les données manipulées et analysée dans le cadre du projet avec la Banque de France ainsi que les données complémentaires collectées pour le projet avec Generali sont quant à elles publiques.

En ce qui concerne la conformité au règlement général sur la protection des données (RGPD), l'ensemble des données que j'ai traité au cours du stage était anonyme.

2 Résumé technique

Ce rapport revient sur les 25 semaines de stage que j'ai effectué du 2 septembre au 21 février 2020 au sein du Datalab de l'Institut Louis Bachelier. Après avoir présenté l'Institut Louis Bachelier l'activité du Datalab, ce rapport présentera mes contributions aux différents projets ayant été menés sur cette période. Ce rapport présentera notamment le travail réalisé sur deux projets. Dans un premier temps seront détaillées mes contributions sur un projet de recherche en partenariat avec la Banque de France visant à identifier et caractériser les états sous-jacents de l'économie avec de la classification automatique par apprentissage non supervisé et dans un second temps celles sur un projet commissionné par le groupe Generali cherchant à exploiter les données libres d'accès et en particulier les avis et commentaires pour modéliser les risques des petites et moyennes entreprises. Ce rapport évoquera les techniques de machine learning et de traitement du langage naturel employées au cours de ces projets et en présentera certains résultats. Ce rapport présentera dans un dernier temps mes conclusions et mon ressenti vis-à-vis de ce stage.

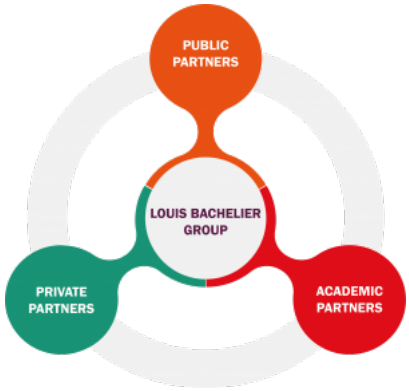
3 Présentation de l'Institut Louis Bachelier et de son Datalab

3.1 L'Institut Louis Bachelier

Le groupe Louis Bachelier a pour vocation de financer, développer et promouvoir la recherche d'excellence en économie et finance. Le groupe Louis Bachelier regroupe la Fondation du Risque (FDR), l'Institut Europlace de la Finance (IEF) et enfin l'Institut Louis Bachelier (ILB) au sein duquel j'ai effectué ce stage. L'Institut Louis Bachelier, basé au Palais Brongniart, a été créé en 2008 par la Direction Générale du Trésor (DGT) et la Caisse des Dépôts et Consignations (CDC) pour stimuler les échanges de savoir et les collaborations transversales dans le but de favoriser l'émergence d'une économie et d'une finance durables. Il est le moteur d'un réseau de recherche partenariale en économie et finance. Le but de cet organisme est de favoriser le financement, la diffusion et la valorisation de la recherche avec les entreprises des secteurs Banque, Finance et Assurance. L'ILB met en relation des entreprises qui veulent initier des programmes de recherche avec des laboratoires et leurs équipes de chercheurs. À ce jour, le réseau de l'ILB regroupe :

- 39 institutions académiques (Polytechnique, HEC, Dauphine, ENSAE, TSE, etc.),
- 85 entreprises privées (Société Générale, Crédit Agricole, Air Liquide, Total, etc.),
- des institutions publiques (Pôle Emploi, différents ministères, etc.).

Cela lui permet de proposer, à ses partenaires privés ou publics, les chercheurs les plus adaptés à leurs besoins et, aux académiques, des financements privés de qualité. L'ILB compte actuellement plus de 50 programmes de recherche dits Chaire ou Initiative de Recherche ayant mobilisé plus de 500 chercheurs. Ces programmes de recherche sont regroupés autour de quatre sujets traitant des transitions majeures en cours : environnementale, démographique, financière et numérique. L'ILB se veut être un institut effectuant la transition de la recherche vers l'entreprise, au service de l'innovation.



3.2 Le DataLab

En complément des programmes de recherche de long terme, l'ILB propose aux entreprises, chercheurs et pouvoirs publics un appui pour l'analyse et la modélisation statistique à travers un Datalab dont l'intérêt est de recueillir et de travailler sur des projets de recherche appliquée en Data Science avec une concentration particulière sur l'application de modèles de machine learning. Ce Datalab a été créé en 2015 et compte actuellement une dizaine d'ingénieurs et de stagiaires encadrés par Louis Boulanger, directeur du Datalab et Jean-Michel Beacco, directeur général de l'ILB et professeur associé en Finance à l'Université Paris-Dauphine. À ce jour, le Datalab réalise des projets entre autres avec le groupe Generali, le Crédit Agricole, la Société Générale et la Banque de France.

4 Présentation des missions

Au cours de mes 25 premières semaines de stage, j'ai été amené à travailler sur quatre projets différents. Trois de ces projets ont été menés en parallèle. Le premier est un projet en partenariat avec la Banque de France et en particulier Jean Boissinot, conseiller spécial aux gouverneurs à la Banque de France, qui vise à évaluer la stabilité du système économique et financier avec de l'apprentissage non supervisé. Il s'agit de mon projet principal et a représenté au moins 50% de mon temps de travail réparti sur la totalité de la durée de mon stage. Une grande proportion de mon temps de travail a ensuite été affecté à un projet commissionné par le groupe Generali visant à modéliser le risque MRC en s'appuyant sur l'Open Data. J'ai été particulièrement impliqué dans ce projet d'octobre à la clôture du projet à la fin du mois de décembre. J'ai aussi été amené à contribuer à deux projets de moins grande envergure, l'un commissionné par Asment Temara cherchant à modéliser les temps de trajet de véhicules de livraison et l'autre, couvrant mes 5 dernières semaines de stage, cherchant à modéliser le prix de vente de diamants pour la start-up Everdiams.

4.1 Évaluation de la stabilité économique et financière par Machine Learning

4.1.1 Présentation

À ce jour, la surveillance et l'évaluation de la stabilité économique et financière est plus un art qu'une science. Bien que de nombreuses données soient disponibles, celles-ci sont très hétérogènes (fréquence, nature...) et si volumineuses qu'il est difficile d'en extraire du sens. La donnée est aujourd'hui choisie spécifiquement pour illustrer des faits, des évènements ou des états ayant déjà été identifié auparavant à dire d'expert. L'objectif de ce projet est d'employer des méthodes de Machine Learning pour extraire du sens de l'ensemble des données économiques et financières. Le sujet a déjà été approché avec de l'apprentissage supervisé dans un objectif de prédiction de crises. Cet objectif se révèle toutefois trop ambitieux et ne permet pas de capter certains phénomènes sous-jacents déterminants. L'objectif plus réaliste du projet entrepris a donc été de chercher à caractériser les états sous-jacents du système économique et financier en s'appuyant sur de

l'apprentissage non supervisé de manière à pouvoir réinterpréter ces états ex-post en termes de stabilité ou stress.

4.1.2 Contributions

Dans un premier temps, aux côtés d'un autre stagiaire, Adrien Akar et avec les conseils de Driss Lamrani (ILB), Louis Boulanger (ILB) et Jean Boissinot (Banque de France), nous avons procédé au prétraitement des données mises à notre disposition. Il s'agit d'un ensemble de 340 variables concernant essentiellement des indicateurs économiques ou des valeurs de marché financier. Dans la suite, j'ai été en charge de la modélisation par apprentissage non supervisé des états du système économique et financier à partir de ces données avec Driss Lamrani en appui pour l'interprétation ex-post des résultats. Au-delà de la modélisation par apprentissage non supervisé, j'ai exploré et implémenté des méthodes de "Consensus Clustering" pour améliorer la robustesse des résultats. Dans un dernier temps, avec les conseils de Louis Boulanger, j'ai procédé à une analyse complète de cette modélisation en se focalisant particulièrement sur la robustesse et l'interprétabilité ex-post des résultats.

En lien avec ce projet, j'ai fait une étude et une présentation sur le sujet du Consensus Clustering. Ces présentations d'équipe sont une initiative du Datalab pour se former et se tenir à jour des dernières avancées dans le domaine du machine learning. Comme chaque stagiaire, j'ai eu à présenter la théorie et une implémentation d'un sujet de machine learning pendant 15 à 30 minutes devant l'ensemble de l'équipe sur un sujet de machine learning: le Consensus Clustering.

4.2 Modélisation du risque des PME avec l'Open Data

4.2.1 Présentation

La compagnie d'assurance Generali souhaite exploiter l'Open Data pour enrichir sa modélisation du risque des PME. Le projet entrepris par l'ILB s'inscrit dans cet objectif et se concentre sur la modélisation de la sinistralité du portefeuille d'assurances multirisques commerce (MRC) de Generali et plus spécifiquement des hôtels, bar et restaurants. L'objectif est de collecter une variété de données libres d'accès et d'en évaluer la pertinence pour la prédiction des sinistres de ces établissements. Les données ciblées dans le

cadre de ce projet sont tout d'abord les avis laissés par les clients consommateurs sur la plateforme TripAdvisor dans un second temps les statistiques géographiques et données publiques de l'INSEE et du gouvernement. Ce projet implique d'une part le développement d'un module de collecte des données et plus spécifiquement d'un module de web-scraping et dans un second temps d'une analyse de ces données passant par des méthodes de traitement du langage naturel (Natural Language Processing, NLP) permettant d'évaluer leur influence dans la prédiction des sinistres.

4.2.2 Contributions

J'ai rejoint le projet à trois mois de son terme. À ce stade, le module de web-scraping permettant de collecter les données issues de la plateforme TripAdvisor était terminé. Ma tâche dans un premier temps a été d'identifier les statistiques et les ensembles de données potentiellement pertinents mises à disposition par le gouvernement et l'INSEE et d'effectuer les traitements nécessaires à leur intégration au modèle de prédiction. Dans un second temps, au côté de Melchior Savigneux (ILB), j'ai participé au plongement lexical ("embedding") des commentaires laissés par les clients sur la plateforme TripAdvisor et l'évaluation de leur apport à la modélisation des risques MRC. Dans un dernier temps, j'ai participé à l'évaluation de la robustesse de nos résultats clôturant le projet.

4.3 Autres missions

4.3.1 Estimation des temps de trajets de véhicules de livraison

4.3.1.1 Présentation et contributions

Asment Temara, branche marocaine du groupe cimentier Votorantim Cimentos cherche à estimer les durées des livraisons de leurs camions. Pour ce faire, un dispositif de collecte de données a été installé sur une partie de leur flotte de véhicules relevant des coordonnées GPS horodatées au cours de leurs trajets. À partir de ce grand volume de données GPS et d'un historique des activités de l'entreprise, j'ai proposé et implémenté une méthode impliquant de l'apprentissage non supervisé pour construction une base d'observations des livraisons et itinéraires distincts empruntés par ces camions.

4.3.2 Prédiction de prix de vente de diamants

4.3.2.1 Présentation et contributions

Everdiams est une start-up fondée en 2019 cherchant à créer un fonds négocié en bourse (Exchange Traded Fund, ETF) basé sur le diamant. Dans cet objectif, Everdiams souhaiterait automatiser la valorisation des diamants qui à ce jour est exclusivement faite à dire d'expert. Un extrait d'une base de données des transactions de diamants de l'IDEX (International Diamond Exchange) a été mis à notre disposition. Ces données décrivent les caractéristiques physiques principales des diamants telles que le carat, la coupe, les dimensions, la couleur et autres ainsi que le prix de vente. Nous avons donc procédé à une analyse exploratoire de ces données afin de construire un modèle de régression pour la prédiction de ces prix.

4.4 Outils et technologies

Les travaux sur l'ensemble de ces projets ont été développés presque exclusivement en Python dans un environnement Linux (Ubuntu). L'essentiel des données traitées se présente sous forme d'ensemble de fichiers (CSV ou tableur). Les bibliothèques Numpy et Pandas ont été largement utilisées pour manipuler ces données. Les tâches d'apprentissage ont été développées en s'appuyant sur les modèles de la librairie scikit-learn ainsi que des bibliothèques spécifiques comme hmmlearn pour certains modèles markoviens et Cluster_Ensembles pour l'implémentation du "Consensus Clustering". Enfin, les représentations graphiques et visualisations des données ont été essentiellement réalisées en s'appuyant les bibliothèques matplotlib et Seaborn.

4.5 Prise de recul

Le projet d'évaluation de la stabilité économique et financière par Machine Learning en partenariat avec la Banque de France se distingue des autres projets sur lesquels j'ai pu travailler. Il s'agit du projet sur lequel je me suis le plus investi et là où les projets commissionnés par le groupe Generali et Asment Temara s'apparentent à des projets de conseil avec un planning et des objectifs bien définis, celui-ci s'apparente plus à un projet de recherche, au sujet large et dont les objectifs évoluent et se diversifient en fonction des avancées faites. Mes contributions sur ce projet ont révélé qu'une approche non supervisée est prometteuse et permet une bonne interprétation des états

sous-jacents du système économique et financier. Ces travaux s'inscrivent tout de même dans une étude où d'autres pistes restent à explorer.

Mes contributions aux projets commissionnés par le groupe Generali, Asment Temara et Everdiams ont représenté l'autre moitié de mon temps de travail. Ces travaux se sont chacun inscrit dans un livrable final remis au client. Dans le cas du projet d'exploitation de données libres pour la modélisation des risques des PME, il s'agissait d'une étude exploratoire d'un sujet. Nos travaux ont permis de révéler et d'exclure différentes approches pour la modélisation des sinistres d'un portefeuille d'assurance MRC. En ce qui concerne le projet d'estimation des temps de livraison, la qualité des données n'était pas suffisante pour produire un rendu aussi sophistiqué que je l'espérais. Une amélioration du processus de collecte des données chez le client permettrait de développer quelque chose de plus intéressant et innovant. Enfin, en ce qui concerne le projet Everdiams, la propreté des données initiales nous a permis de présenter une analyse et un modèle de prédiction pertinent et tout à fait satisfaisant.

5 Réalisations

Dans cette partie je présenterai les travaux entrepris et mes contributions sur les deux projets principaux de mon stage. Dans un premier, je détaillerais le contexte, l'approche, mes réalisations et les résultats obtenus sur le projet de recherche entrepris aux côtés de la Banque de France sur la caractérisation des états du système économique et financier. Je ferai ensuite de même pour le projet de modélisation des risques des PME commissionné par le groupe Generali.

5.1 Évaluation de la stabilité économique et financière par Machine Learning

5.1.1 Contexte et problématique

À ce jour, on ne dispose pas d'un modèle standardisé permettant de surveiller et d'évaluer la stabilité du système économique et financier. Bien que ce soit une tâche essentielle pour définir et décider de politiques économiques, la caractérisation de la stabilité du système économique et financier est souvent assez subjective, elle s'apparente plus à un art qu'une science. Une grande quantité et diversité de données économiques et financières sont aujourd'hui disponibles. Ces données sont toutefois très hétérogènes. Même lorsque les données traitent d'un même thème, ces données présentent entre autres des différences en fréquence temporelle, d'échelle et d'unité. Les données sont de plus très volumineuses au point qu'il est difficile d'extraire du sens de l'ensemble de celles-ci. Les experts se contentent aujourd'hui d'illustrer des tendances ou des faits observés et pré-identifiés avec une sélection limitée bien précise de ces données.

Le machine learning semble être une solution potentielle pour extraire du sens de ces ensembles d'informations. Une première approche, poursuivie par Rey et Fouilliard ou encore la Bundesbank, a été de développer un modèle de prédiction de crise systémique. Cette approche semble toutefois être trop ambitieuse. La capacité d'interprétation du modèle est en effet mise de côté au profit des performances de prédiction sur un historique donné. Il est difficile de savoir si le modèle est en réalité capable de relever les déséquilibres macro-économiques sous-jacents qui définissent l'état de stabilité du système économique et financier. L'approche plus réaliste que nous adoptons dans ce

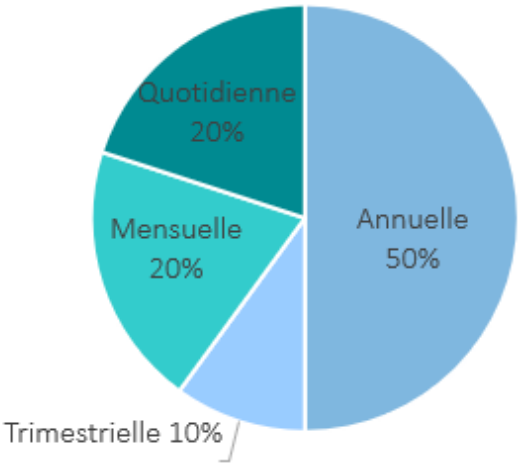


Figure 1: Proportion des données originales par fréquence temporelle

projet est de développer un modèle capable avant tout d'identifier et de caractériser les états sous-jacents du système économique et financier en matière de stabilité ou de stress. En partant d'environ 350 variables décrivant divers aspects du système économique et financier, nous avons proposé une approche non supervisée de classification automatique pour reconnaître ces états.

5.1.2 Les données : présentation et prétraitements

Un ensemble de données publiques a été mis à notre disposition par la Banque de France, il s'agit de 340 séries temporelles décrivant 149 caractéristiques distinctes du système économique et financier. Ces séries temporelles couvrent la période 2000 à 2019. Une exploration rapide de ces données a permis de révéler qu'il existait au sein de ces données des différences de fréquence temporelle significatives et un grand nombre de valeurs manquantes.

Les prétraitements de ces données s'effectuent en plusieurs étapes. Les variables de fréquence temporelle différentes sont traitées indépendamment. Après standardisation des unités à travers chaque ensemble de variables, on identifie et on met de côté les variables peu renseignées, présentant moins de trois observations. Les valeurs manquantes restantes sont ensuite interpolées en propageant la valeur précédente. Pour mieux prendre en compte la dimension temporelle de nos données, nous avons calculé, pour certains sous ensembles de variables identifiés à dire d'expert par Driss Lamrani, la

différence relative ou bien le log-ratio entre deux observations consécutives. On effectue ensuite une analyse par composante principale suite à laquelle on conserve le sous-ensemble de composantes principales minimal permettant de retenir au moins 90% de l'inertie totale du jeu de données. Dans un dernier temps, les données quotidiennes sont agrégées en calculant la moyenne et l'écart-type mensuel. Ce prétraitement est encapsulé dans un module python en sortie duquel quatre ensembles de données sont produit sous forme de fichier CSV.

- Les données quotidiennes :
  - Période de Janvier 2007 à Juin 2019
  - Valeurs quotidiennes agrégées au mois (moyenne et écart type)
  - 150 observations de 106 variables (2 * 53 variables)
- Les données mensuelles :
  - Période d'Avril 2008 à Novembre 2016
  - 104 observations de 45 variables
- Les données trimestrielles :
  - Période de Juin 2008 à Décembre 2016
  - 35 observations de 16 variables
- Les données annuelles :
  - Période de 2008 à 2016
  - 9 observations de 102 variables

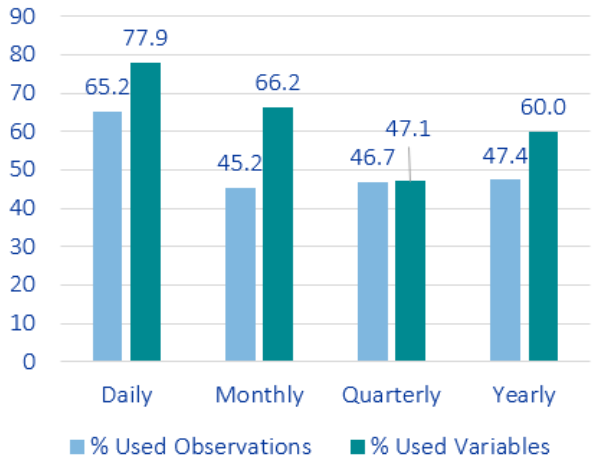


Figure 2: Proportion des observations et variables originales retenues par fréquence temporelle

5.1.3 Apprentissage non supervisé, classification automatique

Dans la suite de l'étude, nous nous sommes focalisés sur les données relevées à la fréquence quotidienne. Les variables en question sont essentiellement des prix sur le marché financier. L'ensemble de données correspondant, 150 observations des moyennes et écart-types mensuelles de 53 prix de produits financiers distincts, soit 106 variables au total. Cet ensemble de données est centré et réduit avant d'être soumis à trois algorithmes de classification automatique non supervisée de nature diverse:

- Kmeans

  L'algorithme des Kmeans consiste en un algorithme d'optimisation cherchant à minimiser l'inertie intra-classe ou de manière équivalente à maximiser l'inertie interclasse du nuage d'observations. Cet algorithme a été implémenté pour se donner une première référence de classification automatique non supervisée de nos données. L'algorithme est particulièrement performant pour identifier des classes relativement bien regroupées dans l'espace et de dimension comparables.

- Gaussian Mixture Model

  Le Gaussian Mixture Model s'appuie sur un algorithme d'espérance-maximisation (EM) pour trouver, d'après les données, les paramètres

optimaux d'un mélange de lois Gaussiennes. Il s'adapte, contrairement au k-means, a des classes de formes et tailles plus diverses. Toutefois, comme le k-means, il n'est pas adapté aux cas de classes non convexes. Ce modèle affecte à chaque point une probabilité d'appartenance à chaque classe et permet ainsi de raffiner les conclusions qu'on en tire. Dans notre cas, cela nous peut nous permettre d'identifier des moments d'incertitude ou de transition l'état du système économique et financier.

- Hidden Markov Model

  Le Hidden Markov Model (HMM) suppose que les observations d'un jeu de données sont régies par une chaîne de Markov dont les états sont à priori inconnus ("cachés"). Le modèle s'appuie sur un algorithme "forward/backward" suivi d'un algorithme EM pour trouver les paramètres optimaux de la loi de probabilité jointe du système. L'algorithme de Viterbi permet enfin d'associer un état à chaque observation de manière à maximiser cette probabilité jointe. L'intérêt du HMM dans notre cas est le fait qu'il permet de prendre en compte la nature temporelle de nos données. Au vu de la nature de nos données, on peut s'attendre à ce qu'il y ait une certaine dépendance temporelle non négligeable. Une telle modélisation markovienne des états sous-jacents du système économique et financier est d'autant plus pertinente qu'on est capable d'estimer les probabilités de transition d'un état à un autre.

L'évaluation des résultats de classification automatique non supervisée est un problème non trivial. Contrairement au cas supervisé, on ne dispose pas d'une référence d'après laquelle on peut calculer un score ou une mesure de la qualité de nos résultats.

L'évaluation de nos résultats s'est appuyé d'une part sur des mesures relatives à l'inertie intra-classe et interclasse comme l'index de Calinski-Harabasz. Nous avons aussi, à titre plus indicatif, exploré nos résultats avec des méthodes de visualisation avancées, en particulier le T-distributed stochastic neighbour embedding (T-SNE). Cela dit, l'évaluation de la qualité de nos résultats repose essentiellement sur la capacité d'interprétation de ceux-ci à dire d'expert. Plus spécifiquement, on cherche à obtenir des classes relativement bien regroupées dans le temps de manière à ce qu'un expert puisse associer certaines classes à des périodes bien définies associées à tel ou tel évènement ou fait observé. Les paramètres finaux de nos modèles sont

donc le résultat d'une optimisation à la fois de certains critères spécifiques aux modèles, mais surtout de l'interprétabilité des états identifiés.

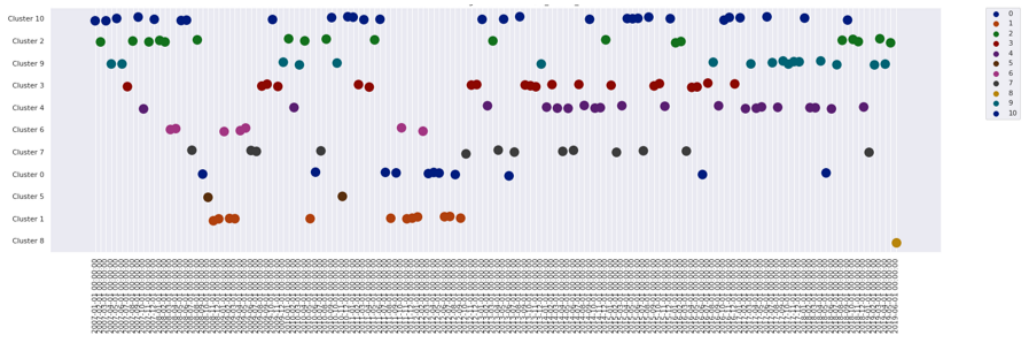


Figure 3: Représentation dans le temps des classes résultantes d'une classification automatique par GMM

5.1.4 Interprétation et caractérisation des résultats

L'interprétabilité des états identifiés est le point fondamental de notre étude. Pour procéder à la caractérisation des états résultant de notre classification automatique, nous avons dans un premier temps regroupé manuellement les 53 variables en 6 catégories de produits financiers différents (Credit Default Swaps, Obligations Souveraines...). Pour chaque catégorie, nous avons retenu la première composante principale d'une Analyse par Composante Principale effectuée sur les variables de la catégorie en question. Au minimum 70% et en médiane plus de 80% de l'inertie totale est retenue pour chaque catégorie.

On se propose ensuite d'évaluer à quel point chacune de ces catégories de variable est discriminante dans nos résultats de classification automatique. Pour ce faire, on entraîne un modèle de classification supervisé : une forêt aléatoire, sur ces données catégorisées et classifiées. Au vu de l'algorithme de développement des arbres de décision composant les forêts aléatoires, les profondeurs de chaque variable dans ces arbres apportent de l'information sur l'importance discriminatoire de ces variables. L'implémentation du classifieur par forêt aléatoire de la bibliothèque scikit-learn permet de collecter une mesure de ces importances discriminatoires.

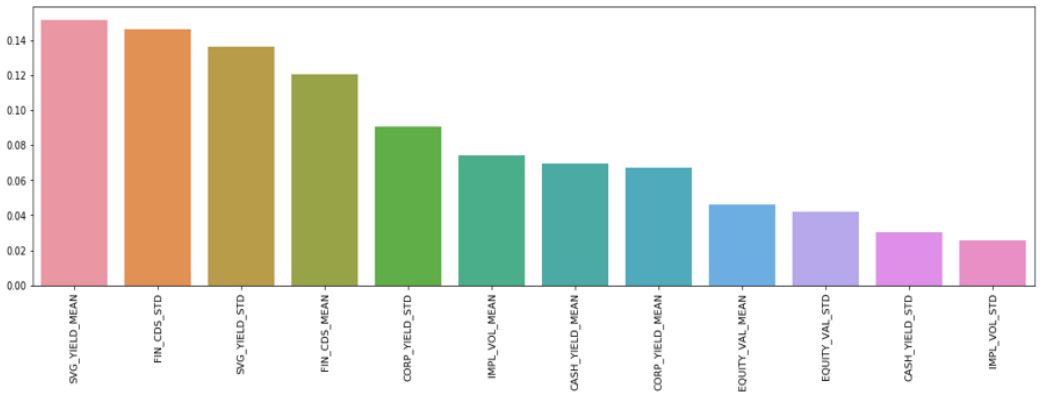


Figure 4: Importance discriminatoires des variables dans la forêt aléatoire de classification

Dans l'objectif de comprendre plus spécifiquement les différences entre les classes résultantes, nous avons procédé à une série de tests statistiques. Nous avons cherché à identifier des différences de distribution des catégories de variables entre les classes. Le volume des données le permettant, nous avons effectué des tests de différence de moyenne et mesuré la distance de Wasserstein entre toutes les combinaisons de classes possibles.

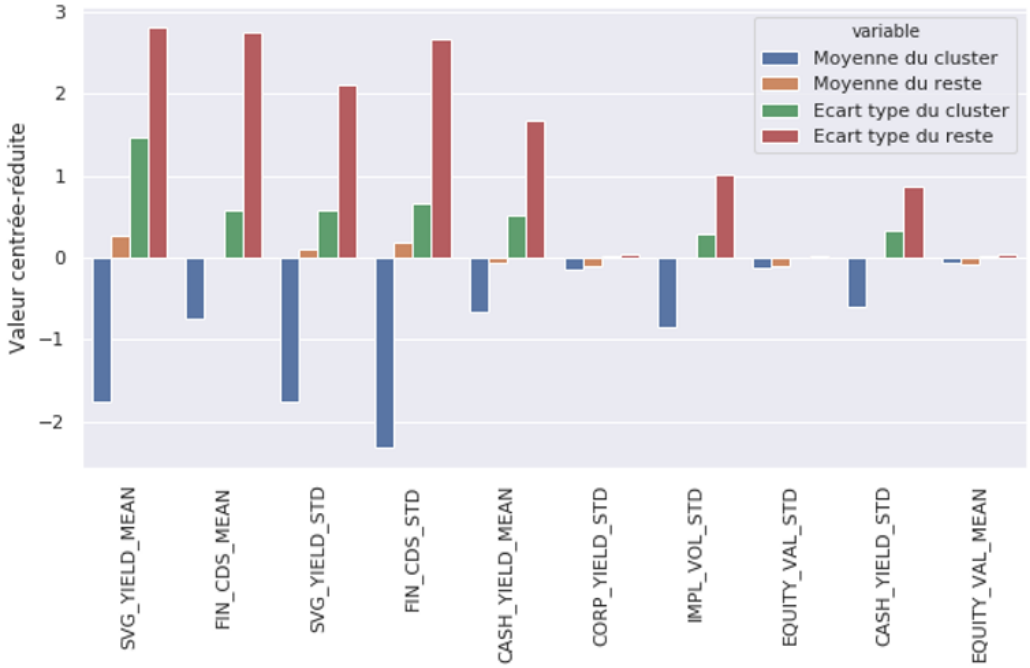


Figure 5: Moyenne et écart-type des variables significatives au niveau 0.05 pour un test d'égalité des moyennes de Welch pour une classe donnée contre le reste

Dans un dernier temps, nous nous sommes intéressé aux transitions entre classes dans le temps. Nous avons calculé pour ceci les probabilités à priori de transition entre chaque paire de classe. Pour chaque classe on calcule le ratio du nombre de transition vers chaque classe sur le nombre total de transitions observées (une transition pouvant aussi être d'une classe vers elle-même).

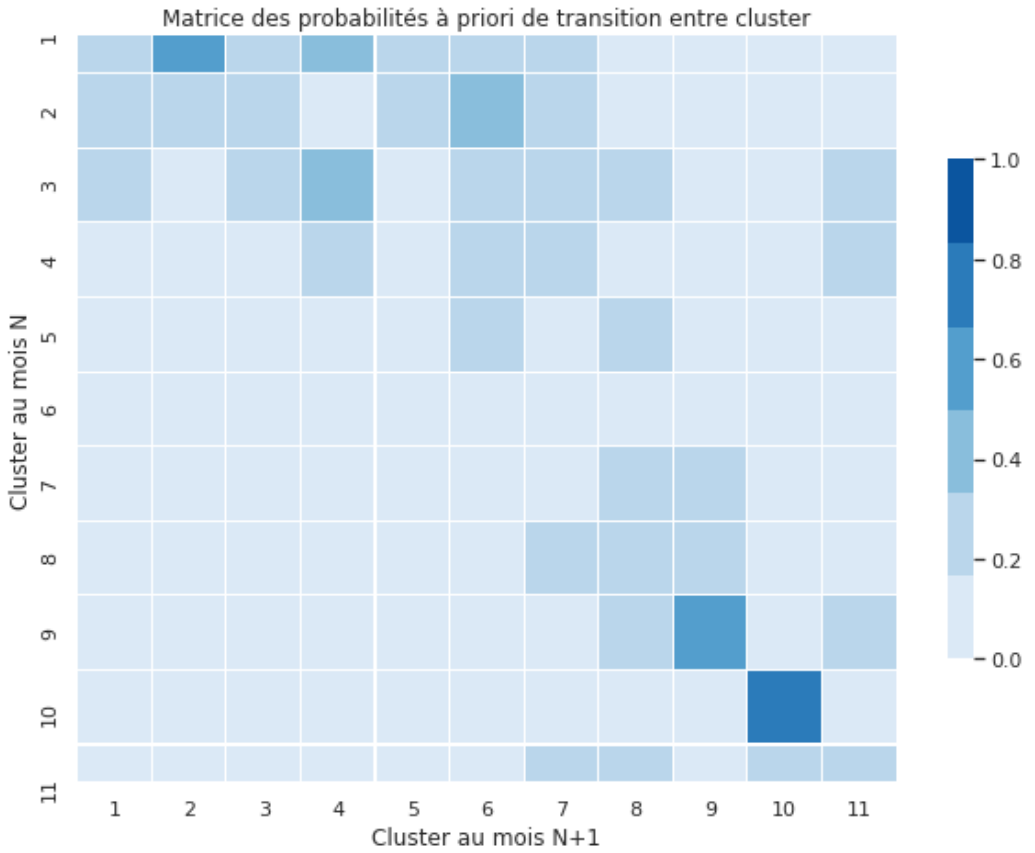


Figure 6: Matrice des probabilités à priori de transition entre chaque paire de classe

Ces démarches de caractérisation des classes résultantes ont permis d'associer à celles-ci une dénomination reflétant les états réels de tension ou de détente spécifique du système financier.

5.1.5 Consensus Clustering

Dans le cadre de cette étude, nous avons souhaité explorer des techniques de Concensus Clustering (ou Ensemble Clustering) qui consistent, à trouver la classification (clustering) représentant le plus fidèlement un ensemble de classifications différentes sans avoir recours ni aux données initiales ni aux algorithmes de classifications sous-jacents. Le consensus clustering est particulièrement intéressant pour plusieurs raisons :

- Le recyclage de connaissance

  Le consensus clustering permet d'exploiter des résultats classifications traitant d'un même sujet mais dont les variables sont potentiellement différentes. Dans notre cas, une approche par consensus clustering pourrait permettre d'intégrer des résultats de classifications résultantes de variables à fréquences temporelles différentes.

- La robustesse

  Le consensus clustering est, de manière analogue au cas supervisé, une approche d'apprentissage par ensemble qui permet d'obtenir des résultats plus robustes. On peut en effet employer des méthodes de "bagging" et par consensus clustering trouver une classification moins sensible à de petits changements dans l'ensemble de données initial.

- Les systèmes distribuées

  Le consensus clustering peut s'avèrer d'une utilité particulière dans les systèmes distribués, il offre un moyen d'effectuer une classification automatique distribuée en traitant tout d'abord les données "in situ" avant d'en centraliser un résultat final. Ceci peut s'avérer particulièrement utile lorsque les données sont trop volumineuses pour être traitées de manière centralisée ou bien lorsque les données sont soumises à certaines contraintes de confidentialité. À ce jour, ce n'est pas le cas dans notre étude puisque l'on travaille avec un volume restreint de données publiques, mais ce pourrait le devenir dans le futur.

Le consensus clustering se formalise comme un problème d'optimisation dans lequel on cherche la classification maximisant l'information mutuelle moyenne partagée avec les classifications données en entrée. Au stade actuel de l'étude, le consensus clustering a été implémenté à titre d'essai pour identifier la classification maximisant l'information mutuelle entre les trois méthodes de classification automatiques employées (Kmeans, GMM et HMM).

Le consensus clustering a fait l'objet de ma présentation d'équipe. J'ai en effet réalisé un exposé des concepts théoriques, des cas d'usage et de l'implémentation du consensus clustering pendant une demi-heure devant une dizaine de membres du Datalab.

5.1.6 Évaluation de la robustesse des résultats

Nous avons développé plusieurs procédures pour évaluer la robustesse de nos données. Dans un premier temps, nous cherchons à évaluer la robustesse de la procédure de classification générale de bout en bout. Pour ce faire, nous avons choisi une approche de bootstrap qui consiste à tirer des échantillons de 80% des données initiales à la procédure, à les soumettre à la procédure de classification automatique et de mesurer l'information mutuelle partagée par la classification résultante avec la classification automatique réalisée sur la totalité des données pour les points correspondants (Firgure 6).

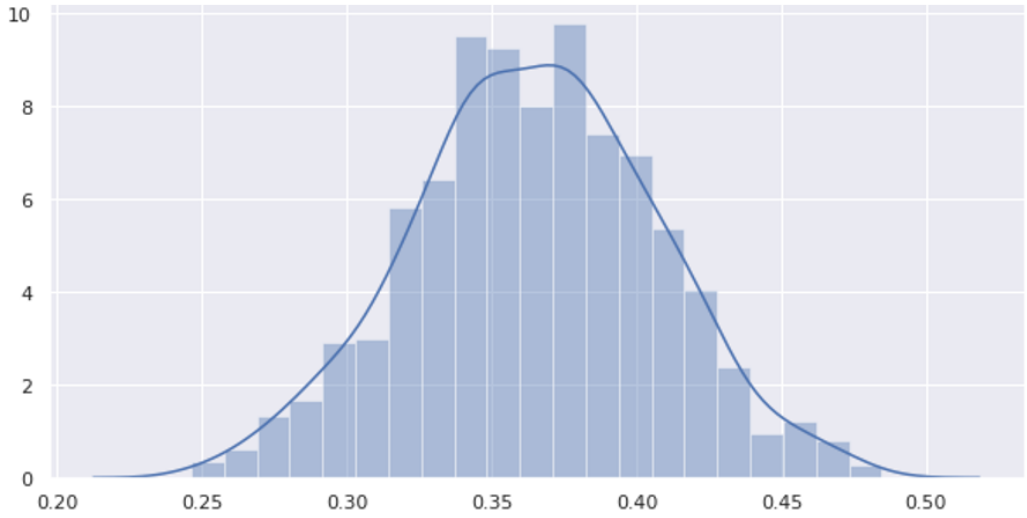


Figure 7: Distribution de l'information mutuelle normalisée pour 1000 échantillons de 80% des données

Dans un second temps, on s'est intéressé à la robustesse de la procédure dans le temps. Nous avons choisi de mesurer l'information mutuelle partagée par une classification automatique réalisée sur l'ensemble des observations faites avant une date limite donnée et la classification réalisée sur la totalité des données. Nous avons mesuré cette information mutuelle pour une date limite variant avec un pas d'un mois le long de l'intervalle de temps couvert par nos données. Comme on pouvait s'y attendre on observe une tendance croissante de l'information mutuelle proportionnellement à l'intervalle de temps considéré (Firgure 7).

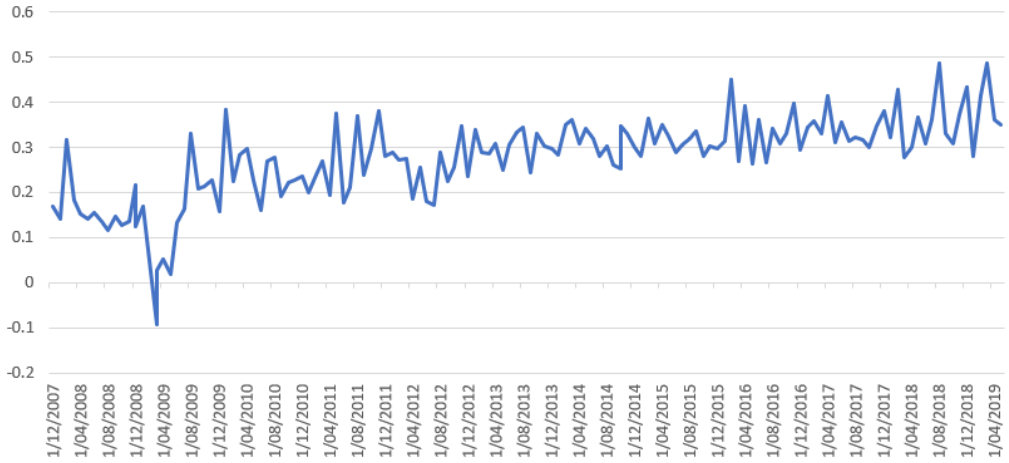


Figure 8: Information mutuelle normalisée partagée avec la classification sur la totalité des donnée et la classification des données avant une date limite donnée en fonction de la date limite conisdérée

Pour affiner cette évaluation de la robustesse dans le temps, nous avons réalisé la même procédure mais en s'intéressant à l'information mutuelle partagée une classification réalisée sur les observations faites avant une date limite donnée et les observations faites avant le mois directement suivant cette date limite. On observe ainsi que l'information mutuelle partagée par deux classifications à un mois d'intervalle est en moyenne assez constante mais décroissante en matière de variance (Firgure 8).

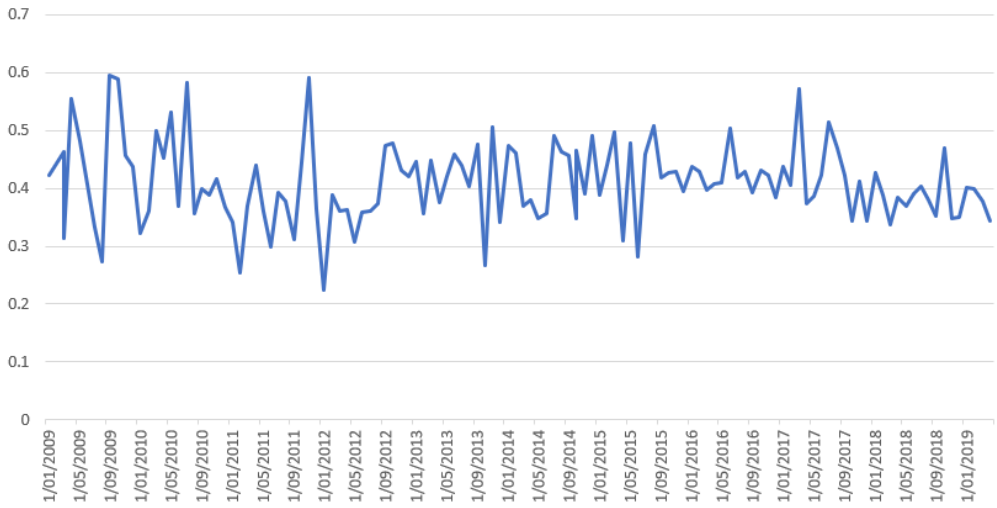


Figure 9: Information mutuelle normalisée partagée par deux classifications réalisée sur les données antérieures à une date limite et le mois directement suivant cette date limite

5.1.7 Conclusions et perspectives futures

Le travail effectué a révélé qu'une approche par apprentissage non supervisé est prometteuse pour l'identification et la caractérisation des états du système économique et financier. On est en effet capable avec des méthodes simples de retrouver des classes interprétables ex-post, reflétant des évènements réels et relativement robustes Nous avons travaillé sur un sous-ensemble de données publiques assez restreint, à la fois en termes de sujet puisqu'il s'agit presque exclusivement de valeurs de produits financiers, mais aussi en profondeur temporelle. Il est prévu dans la suite de l'étude d'intégrer des données relatives à d'autres thématiques comme, entre-autres, des prix de commodités, des taux de change et des indices macro-économiques ainsi que d'étendre la période temporelle étudiée. Il serait intéressant à ce stade du projet d'approfondir individuellement la modélisation par GMM en s'appuyant sur les degrés d'appartenances aux classes associés à chaque observation pour identifier les périodes d'incertitude et de transition du système économique et financier. Il serait d'un autre côté intéressant de creuser plus loin la modélisation par HMM afin d'obtenir une bonne représentation des probabilités de transitions d'un état à un autre. Enfin, dans le cadre de ce projet, un groupe d'étudiant de l'ENSAE a été chargé d'implémenter une

méthode de classification automatique par factorisation de matrice sparse. Cette approche est innovante, une étude comparative des résultats entre celles-ci et celle que nous avons implémentée pourrait mener à des conclusions intéressantes.

5.2 Modélisation du risque des PME avec l'Open Data

5.2.1 Objectifs et démarche

La mission « Open data for SMEs risk modelling » porte sur la modélisation du risque du portefeuille MRC (Multi Risque Commerce) des hôtels, bars et restaurants assurés par Generali. Plus particulièrement, il s'agit d'explorer l'apport de l'open data à la modélisation des fréquences et des coûts moyens par type de sinistre du portefeuille MRC.

Il s'agit dans un premier temps d'identifier les sources de données potentiellement pertinentes et qui peuvent être associées à un établissement de la base interne. Une fois ces données identifiées il a fallu les collecter avant enfin d'analyser l'apport de celles-ci à la modélisation des sinistres de ces établissements.

À mon arrivée sur le projet, trois mois avant la fin de celui-ci, le prétraitement et l'analyse des données du portefeuille MRC fourni par GENERALI était terminé. Le choix avait été fait de se focaliser dans un premier temps sur une source de données : la plateforme de revues, d'avis et de conseils touristiques TripAdvisor. Ma tâche a donc été d'une part de participer au traitement de ces données (NLP) et à l'analyse de celles-ci de manière à évaluer leur apport à la modélisation des risques du portefeuille MRC de GENERALI. Dans un deuxième temps, j'ai été responsable de la collecte, de l'intégration et de l'analyse de l'apport de données publiques (INSEE et data.gouv) à cette modélisation. Pour des raisons de confidentialité, je ne ferai pas de présentation des données de la base MRC. Je présenterai toutefois rapidement la collecte et les données extraites de la plateforme TripAdvisor.

5.2.2 Les données TripAdvisor

Afin de collecter un ensemble de données potentiellement pertinentes de la plateforme web TripAdvisor, un module de "web scraping" a été développé, en grande partie par Yating DENG, également stagiaire UTC au Datalab de l'ILB. Il s'agit d'un script python exploitant les libraires « Scrapy » et « Se-

lenium ». La libraire Selenium permet d'automatiser la navigation sur les pages web (simulation de click et autres actions), et la librairie Scrapy permet de récupérer les données contenues dans des balises de pages HTML. Ce module de "web scraping" collecte les données correspondant à l'établissement donné du portefeuille MRC en recherchant la page TripAdvisor correspondante par le biais d'un moteur de recherche ou bien par recherche directe sur la plateforme.

Les données récoltées correspondent aux informations disponibles sur le site TripAdvisor, à savoir pour chaque établissement :

	Restaurants	Hôtels
Volumétrie : données complètes avec informations et commentaires TripAdvisor	8784	1341
Variables		
Nombre d'avis	Oui	Oui
Notes et Proportions(Excellent/Médiocre)	Oui	Oui
Horaires travaillées (pourcentage hebdomadaire, travail du soir et week-ends)	Oui	Non
Equipements : facilité d'accès, service d'alcool et service à table	Oui	Non
Equipements : blanchisserie, espace fumeur	Non	Oui
Fourchette de prix	Oui	Non
Classement (indicateur de l'intensité de la concurrence)	Oui	Non
Informations sur les restaurants proches (distance, note et nombre d'avis moyens)	Oui	Non

Figure 10: Description des variables issues du scraping de la plateform TripAdvisor

Ces données ont été récupérées pour l'ensemble des 10 125 établissements sélectionnés.

5.2.3 Traitement des données TripAdvisor et NLP

Les données présentées dans la partie précédente ont été traitées de la manière suivante : les valeurs manquantes ont été imputées par la médiane. Les variables Équipements, Horaires travaillées (ouverture le week-end, le soir) sont ensuite binarisées et dans un dernier temps, une variable mesurant le pourcentage hebdomadaire d'ouverture est calculée.

Les commentaires quant à eux ont fait l'objet d'un traitement à part pour la modélisation. Nous avons en effet effectué un travail de traitement du language naturel (NLP) afin de trouver une bonne représentation des informations contenues dans ces commentaires qui soient intégrables avec les données de la base MRC.

5.2.3.1 Analyse de sentiment

Nous nous sommes appuyés sur la librairie "textblob-fr" pour effectuer une analyse de sentiment des commentaires et des titres de ceux-ci. La libraire "textblob-fr" intègre un modèle pré-entraîné permettant d'évaluer d'une part la polarité du commentaire et d'autre part la subjectivité de celui-ci. La polarité du commentaire est un score compris entre -1 et 1 indiquant à quel point le texte reflète une opinion négative (-1) ou positive (+1). La subjectivité de la même manière est un score compris entre 0 et 1 évaluant à quel point le commentaire est objectif (0) ou subjectif (1). Pour chaque établissement, nous avons retenu la moyenne de ces deux scores et avons par la suite construit une variable synthétique représentant le sentiment du commentaire par la polarité pondérée par l'objectivité :

$$sentiment = \text{polarité} * (1 - \text{subjectivité})$$

5.2.3.2 Vectorisation des commentaires

D'autre part, nous avons cherché à extraire de l'information des commentaires en utilisant des méthodes de vectorisation (ou embeddings). Cette approche s'est déroulé en plusieurs étapes :

- Constitution de documents :

  L'unité d'étude est l'adresse assurée, ainsi, nous avons regroupé en un unique document tous les commentaires référant à un même établissement. Pour éviter de prendre en compte des commentaires correspondant à des visites post-sinistre, nous avons écarté les commentaires publiés avant le début de la période d'exposition (2015) lorsque c'est possible, les 10 commentaires (maximum) les plus anciens dans l'autre cas.

- prétraitement des données textes :

  En s'appuyant sur un classifieur pré-entraîné de la bibliothèque Spacy, nous avons identifié la langue des commentaires afin de ne conserver seulement les commentaires en français. Nous avons ensuite retiré ou unifié (remplacement par un unique caractère identifié) les nombres et les caractères spéciaux. Nous avons aussi retiré certains "stop-words" (mot commun considéré peu utile) pré-identifiés lorsque recommandé comme pour l'approche par tf-idf. Nous avons ensuite regroupé les mots ou groupe de mots en unités sémantiques ou "token" en s'appuyant sur

la bibliothèque "nltk". Nous avons enfin ramené ces mots ou groupes de mots résultants à leur racine ou radical soit par "stemming" (retrait des terminaisons les plus répandues) ou bien par "lemmatizing" qui s'appuie sur une base de connaissances des radicaux des mots du dictionnaires comme celle proposée par la bibliothèque "Spacy".

- Vectorisation :
  - Modèles Word2Vec et Doc2Vec :

    Le modèle Word2Vec est un modèle d'encodage des mots par un réseau de neurones artificiel en vecteurs de taille fixe donnée de manière à ce que la proximité des vecteurs dans l'espace vectoriel latent reflète la proximité sémantique des mots qu'ils représentent. Nous avons utilisé un modèle pré-entraîné sur le corpus des articles Wikipédia en langue française. Le modèle Doc2Vec est une extension du modèle Word2Vec permettant de transformer non plus les mots mais un document entier, ici un commentaire ou un ensemble de commentaires, en un vecteur de taille fixe. Ce modèle nous a permis d'associer à chaque établissement un unique vecteur représentatif de celui-ci.

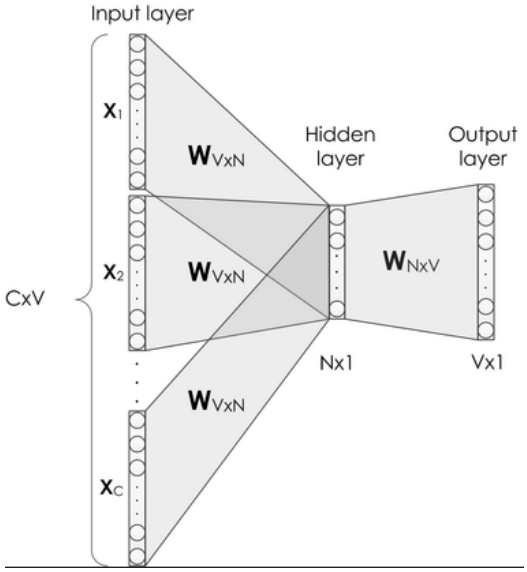


Figure 11: Vectorisation des mots par le modèle Word2Vec

– TF-IDF (Term Frequency Inverse Document Frequency):
Cette approche consiste à représenter les mots par le rapport de leur fréquence d'apparition dans un document sur la fréquence d'apparition de documents présentant ce mot dans le corpus de documents. De cette manière, les documents sont représentés par des vecteurs de la taille du vocabulaire du corpus considéré. La taille de ce vecteur étant ainsi très importante, nous avons retenu un sous-ensemble limité de composantes principales réalisé sur ces vecteurs.

5.2.4 Collecte et prétraitement de données publiques complémentaire

Nous avons étendu la base de données avec des données issues de 17 datasets au total : 11 datasets INSEE, 4 datasets data.gouv, 2 d'autres sources :

- Les datasets INSEE :
  - Données démographiques de 2015 par IRIS (population par tranche d'âge)
  - Revenus déclarés des ménages par unité de consommation - Année 2014 et 2015 par IRIS
  - Base FiLoSoFi 2016 (revenus, pauvreté, niveau de vie des ménages à par Code INSEE
  - Base des équipements touristiques 2019 (nombre d'hôtels, chambres, campings) par Code INSEE
  - Activité des résidents 2015 (chômage, actif, artisans) par IRIS
  - Base des diplômes et formation par IRIS
  - Base des équipements de service et commerce 2018 par IRIS
  - Base des équipements de transport et tourisme (gares, aéroports, hôtels) par IRIS
- Les datasets Data.gouv :
  - Catastrophes naturelles par code postal (nombre et type de périls)
  - Risque sismique par code postal
  - Nombre de zones de sécurité par code postal

- Nombre de crimes et délits par catégorie et par code postal
- Autres sources :
- Base FINESS : nombre d'établissements hospitaliers par code postal
- Base Prométhée : surface incendiée depuis 2015 par IRIS (région méditerranéenne)

La jointure de ces données complémentaires aux données d'origine s'est appuyée exclusivement sur la dimension géographique. Les jointures ont été effectuées au grain géographique renseigné le plus fin possible.

Figure 12: Index géographique sur lequel les données publiques ont pu être jointes aux données de la base MRC

La volumétrie de la base d'étude évolue en fonction des données externes jointes :

- 89% des données d'origine intersectent avec un premier sous ensemble de données INSEE : démographie, activités, diplômes
- On tombe à 45% en ajoutant les données INSEE d'équipement de commerce et de transport
- En considérant toutes les données INSEE, on ne retient plus que 26% des données d'origines
- La base prométhée intersecte exclusivement avec les données relatives à la région méditerranéenne soit 17% des données d'origines

- Pour les données de catastrophes naturelles, on peut conserver entièrement 100% des données d'origine par imputation
- Les autres sources de données individuellement ont une intersection suffisante pour effectuer des analyses (environ 10%) mais insuffisantes lorsqu'on les combine entre elles

Avant de procéder à la modélisation, nous avons effectué quelques prétraitements sur ces données. Nous avons tout d'abord identifié et retiré les variables non pertinentes par nature. Certaines variables ont étés agrégées par catégorie comme les crimes et délits d'un même type dont on retient simplement le compte. Les variables en nombre ont ensuite été retraitées en proportion ou proportion par habitant selon le cas. Nous avons enfin calculé la variation relative et absolue des variables pour lesquelles nous disposions de plusieurs années de données.

5.2.5 Évaluation de l'apport des données à la modélisation

Dans l'étude, les variables cibles sont les fréquences et coûts moyens par type de sinistre.

Dans un premier temps des statistiques descriptives ont été menées. Les corrélations de Pearson et Spearman entre les variables explicatives et les variables cibles n'ont pas donné de résultats significatifs que ce soit pour les données extraites de TripAdvisor ou les données publiques.

Les méthodes de statistiques usuelles n'ayant pas donné de résultat, nous avons cherché à mesurer l'apport des données à la modélisation, par régression, des fréquences et coûts moyens par type de sinistre. Pour cela :

- Nous construisons 2 bases de variables explicatives :
  - Une base dite interne, qui contient uniquement les variables explicatives issues de la base MRC
  - Une base dite globale, qui contient les variables MRC et celles des données complémentaires
- Nous analysons l'apport des variables complémentaires en comparant les régressions des fréquences et coûts moyens faites avec la base interne à celles faites avec la base globale. La comparaison est faite en comparant les scores de performance.

- La métrique de performance est le critère de Gini, en accord avec les pratiques de Generali. Pour rappel, le critère de Gini pour la régression est :

$$Gini = 2 * AUC - 1$$

- Le modèle de régression utilisé ici pour présenter les résultats est le Random Forest, qui va nous servir de base pour détecter du signal. Avant d'utiliser ce modèle de régression, nous avons utilisé des modèles linéaires comme l'ElasticNet, pour lesquels les données TripAdvisor ne permettaient pas d'améliorer la prédiction.

- La modélisation se fait en séparant les données en 85% Train et 15% Test : La base d'entraînement sert à calibrer le modèle et la base de test sert à tester les performances du modèle

En effectuant ce procédé plusieurs fois, il a été remarqué que les scores de performances observés avaient une grande variance d'une combinaison de Train/Test à l'autre (les établissements étant répartis au hasard entre la base d'entraînement et la base de test). La conclusion au sujet de l'apport d'un signal des données TripAdvisor est de ce fait complexifiée. Afin de mettre en place une méthodologie robuste de mesure du signal, il a été décidé de construire des distributions des scores de performances et de les analyser. Pour cela, le processus de régression est réitéré 100 fois pour chaque variable cible et chaque configuration de séparation en Train et Test. Chaque itération fournit une réalisation du score de performance avec et sans variables externes. Le processus global est décrit dans le schéma suivant :

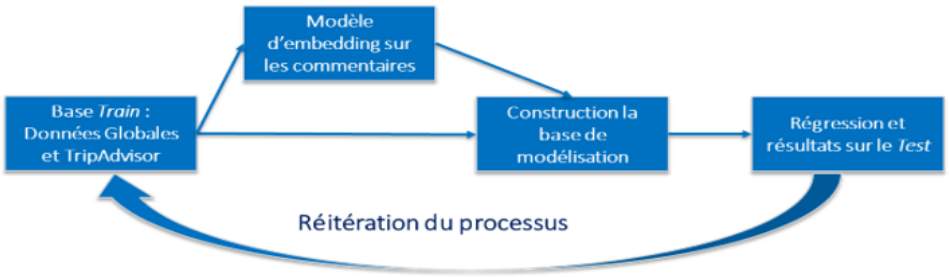


Figure 13: Processus d'analyse de l'apport des variables complémentaires à la modélisation

5.2.6 Résultats et conclusions

La répétition des régressions sur les variables cibles permet d'obtenir plusieurs réalisations des scores de régression avec et sans données complémentaires. De ces réalisations, on compare les distributions des scores, notamment la moyenne, l'écart-type et les quantiles principaux. À partir de ces résultats il est possible de statuer si les variables permettent une meilleure modélisation au sens de la métrique de Gini utilisée. Cependant, cette méthode n'apporte pas d'indice sur la signification du signal qu'apportent les variables complémentaires.

L'analyse des distributions des scores de performance des modélisations de la fréquence est du coût moyen des sinistres montre que les variables issues des données TripAdvisor (données quantitatives et embedding des commentaires) peuvent apporter du signal sur la sinistralité du portefeuille MRC. Cette amélioration est toutefois faible, on observe en effet, dans la meilleure configuration, une amélioration moyenne du score de Gini de 0.5 pour certaines variables cibles seulement. Il est de plus difficile d'interpréter ce signal et en particulier le signal potentiel émis par les embeddings de commentaires.

RESTAURANTS : FREQUENCE RC (TripAdvisor)	Données			Score de performance : Gini					Evolution du Gini : variables internes vs toutes les variables			
	Nombre de variables	Volumétrie Train	Volumétrie Test	Moyenne	Ecart-type	Médiane	quantile 75 %	quantile 25 %	Moyenne	Médiane	quantile 75 %	quantile 25 %
Variables internes et externes	352	7446	1338	**0.18**	0.13	**0.18**	0.25	0.10	0.05	**0.07**	0.16	-0.06
Variables internes	26			**0.13**	0.13	**0.12**	0.23	0.03				

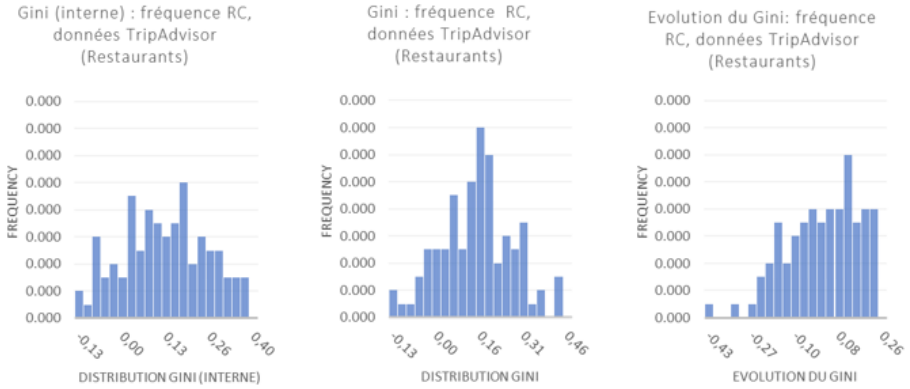


Figure 14: Résultats de l'analyse de l'apport des variables TripAdvisor à la modélisation de la fréquence du sinistre RC

En ce qui concerne les données INSEE, la plupart d'entre elles n'apportent

pas d'information sur le risque MRC et lorsque ces variables semblent avoir du signal, il est difficile de comprendre le phénomène sous-jacent, en effet, on observe parfois que l'ajout de variables de nature incohérente avec le sinistre prédit améliore la moyenne des scores de Gini. Toutefois, les données relatives à la police nationale et la criminalité apportent bien du signal sur le coût moyen d'un vol mais pas sur la fréquence.

FREQUENCE EVENEMENTS CLIMATIQUES (INSEE 26%)	Données			Score de performance : Gini					Evolution du Gini : variables internes vs toutes les variables			
	Nombre de variables	Volumétrie Train	Volumétrie Test	Moyenne	Ecart-type	Médiane	quantile 75 %	quantile 25 %	Moyenne	Médiane	quantile 75 %	quantile 25 %
Variables internes et externes	212	7348	1837	**0,40**	0,07	**0,41**	0,44	0,36	**0,09**	**0,09**	0,13	0,04
Variables internes	26			**0,31**	0,07	**0,31**	0,36	0,26				

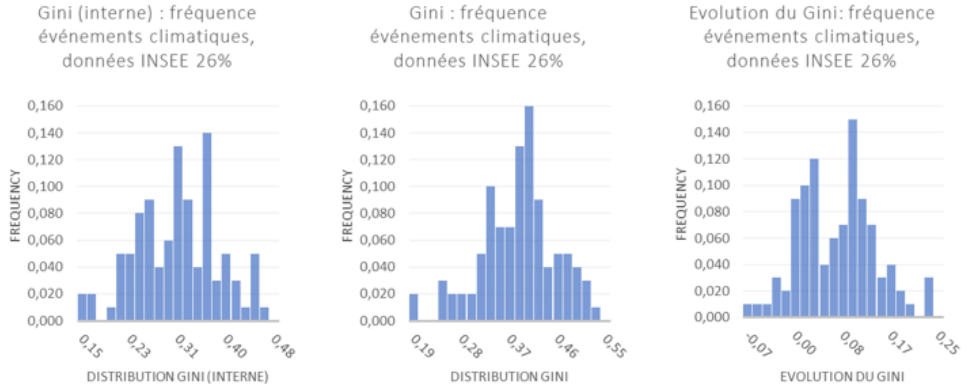


Figure 15: Résultats de l'analyse de l'apport des données de population, infrastructure et revenus des ménages à la modélisation de la fréquence du sinistre liés aux évènements climatiques

Les résultats obtenus à la fin de cette étude semblent donc indiquer que les données complémentaires étudiées ici n'apportent pas ou du moins peu d'améliorations à la modélisation des risques MRC. Les données ne sont toutefois pas sans valeur dans cette étude. Bien qu'elles ne permettent pas, et ce sans grande surprise, de prédire directement la fréquence et le coût moyen des sinistres des clients, elles permettent d'enrichir la base données de Generali et pourront sans doute être exploitées à d'autres fins de modélisation moins ambitieuse comme la simple prédiction d'occurrence d'un sinistre dans le temps.

6 Conclusion

Ce stage au sein du Datalab de l'Institut Louis Bachelier fût une expérience tout à fait bénéfique et valorisante. Je souhaitais, à travers mon stage de fin d'études, m'offrir l'opportunité de monter en compétence en machine learning et me donner l'occasion de travailler sur des projets variés de bout en bout, de la collecte des données jusqu'au rendu final en passant par la préparation des données, l'analyse exploratoire, la modélisation et l'optimisation des paramètres. Mon expérience ici a rempli ces objectifs.

Le premier projet avec la Banque de France m'a permis de mener un travail de Machine Learning dans un contexte de recherche. À travers ce projet j'ai pu monter en compétence en apprentissage non supervisé en découvrant et en implémentant de nouveaux modèles comme le Hidden Markov Model et de nouvelles approches comme le Consensus Clustering. Ce projet m'a d'autant plus permis d'enrichir mes connaissances en macro-économie et en finance.

Le second projet pour le groupe Generali quant à lui m'a permis d'acquérir de nouvelles connaissances et une bonne expérience en traitement du langage naturel (NLP). Ce projet m'a d'autant plus permis d'approfondir mes connaissances des méthodes d'apprentissage supervisé par ensembles et en particulier des méthodes basées sur les arbres de décisions (Random Forests et Gradient Boosting).

L'environnement de travail très ouvert d'esprit et favorisant la collaboration m'a permis de profiter des connaissances de mes collègues et d'en apprendre beaucoup que ce soit en mathématiques appliquées, en machine learning ou en économie et finance.

J'ai identifié toutefois quelques points que je n'ai pas eu l'occasion de traiter ou d'améliorer pendant ce stage. En effet, j'aurais aimé avoir une première expérience professionnelle d'une part en traitement de données à haute volumétrie et d'autre part en deep learning. Ces points me paraissent particulièrement pertinents dans le contexte présent pour tout étudiant aspirant comme moi à devenir data scientist ou ingénieur en machine learning. Une dernière critique que je pourrais faire vis-à-vis de mon expérience ici est un léger manque de principes et bonnes pratiques de développement logiciel, sans doute dû à la jeunesse de l'équipe du Datalab.

Cette expérience reste cela dit très positive et a très bien répondu à mes attentes. Ce stage m'a offert une première expérience solide en tant que data

scientist et confirme ma décision de poursuivre un début de carrière dans cette voie.