Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 21 Septembre 2021

Par : Assan Aziz COULIBALY
Titre : Détéction d'anomalies via l'apprentissage non-supervisé : application à la fraude

Confidentialité : Non

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des Actuaires :
Anthony NAHELOU
Marine HABART
Signature :

Entreprise :
Allianz France
Signature :

Membre présent du jury de l'EURIA :
Brice FRANKE

Directeur de mémoire en entreprise :
Sébastien VITRY
Signature :

Invité :
Jérémy BOULAY
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise :

Sébastien Vitry

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

La charge relative aux indemnisations en arrêt de travail pour l'ensemble des acteurs du marché est en constante évolution ces dernières années, portée à la fois par un effet volume (hausse du nombre de déclarations) et un effet inflation (hausse du nombre de jours déclarés).

Dans un contexte économique marqué par la réduction des coûts, nous nous sommes intéressés aux anomalies constatées sur la base de données relative à la sinistralité en arrêt de travail, base utilisée pour calculer les provisions mathématiques entre autres. Ces anomalies peuvent prendre plusieurs formes (données erronées, aberrantes ou absentes) et sont sources d'erreurs qui peuvent s'avérer coûteuse pour l'entreprise.
Pour arriver à gérer au mieux la sinistralité, il est important d'avoir une réflexion sur les raisons qui peuvent pousser un assuré à se mettre en arrêt, la base de données intiale (issue des systèmes de gestion) a donc été enrichie de diverses variables dites externes disponibles sur le site de l'INSEE.

La détection de ces anomalies a été réalisée via une approche non-supervisée qui a mis en évidence un regroupement d'observations dites "atypiques".
La problématique soulevée par une telle approche est d'évaluer la pertinence des regroupements effectués. En quoi la mise en évidence de ces données anormales présenterait un intérêt ?

Pour répondre à cette problématique, la méthodologie de classification des observations a donc été appliquée à un cas de détection de fraudes dans un cadre supervisé afin d'en évaluer sa pertinence. Ce choix n'est pas anodin en ce sens où une fraude ne correspond pas à une observation dite "normale".
Cette application a démontré le lien qui existe entre observations anormales et fraudes.
Ce lien justifie d'une part la pertinence des regroupements effectués dans la phase d'apprentissage non-supervisé pour détecter les anomalies sur la base de déclaration des arrêts de travail. D'autre part, il prouve l'adéquation des modèles de détection d'anomalies à la problématique de détection des fraudes.

Pour arriver à gérer la complexité du phénomène de la fraude une approche mixte alliant à la fois l'apprentissage supervisé et non-supervisé a été developpée. En effet, il semble évident de supposer que la typologie des fraudes évolue en fonction du temps. Partant de ce constat, avoir une vision combinant un cadre supervisé dont l'objectif serait de détecter les fraudes historiques et une approche non-supervisée axée sur les nouvelles structures de fraudes paraît tout à fait naturel.

Cette étude s'inscrit dans un processus de rationalisation des dépenses en ce sens où les provisions mathématiques étant calculées tête par tête chaque fraude détectée est un gain d'investissement pour l'assureur et une opportunité d'amélioration du ratio de sinistralité.

Mots clefs: Arrêt de travail, Machine Learning, Apprentissage non-supervisé, Apprentissage supervisé, Lien entre Anomalies et Fraudes.

Abstract

Expenditure due to sick leave for all the players in the market have been constantly ´evolving over the last few years. This is due to both a volume effect "increasing in the number of sick leaves" and also an inflation effect " increasing in the number of days spent during sick leave". In an economic context marked by cost reduction we studied the outliers detected in the database of sick leave used to calculate mathematical provisions.

For facing this risk it's important to question oneself about the reasons why an insured worker may take time off , for these reasons we added on the database containing informations on the claims various called external variables available on the INSEE website.

Unsupervised learning is the approach used in our study of outlier detection,this approach helped us to find a cluster of so-called "atypical" observation.
The issue raised by such an approach is to assess the relevance of the cluster. Which additional information did we get from this class of atypical observation ? How could new informations used for ?

For answering to the previous questions, the approach used to detect the outliers was applied to a case of fraud detection through a supervised learning . This choice is not insignificant in the sense that a fraud does not correspond to a so-called "normal" observation.
This application illustrated the link between abnormal observations and fraud.
On the one hand,this link proves the pertinence of the clusters made in the unsupervised learning to detect anomalies in the database of sick leave. On the other hand, it proves that anomaly detection models are well suited to the problem of fraud detection.
In order to deal with the complexity of the fraud phenomenon a mixed approach combining both supervised and unsupervised learning has been developed.
It seems obvious to assume that the fraud structure evolves over time. Based on this observation, having a mixed vision combining a supervised framework whose objective would be to capture historical frauds and an unsupervised framework focused on new fraud structures seems quite natural.

This study is part of a process of optimizing expenditure because each fraud detected is a benefit to the insurance company and an opportunity to improve the loss ratio.

Keywords: Sick leave, Machine Learning, Unsupervised Learning, Supervised Learning, Outliers and Fraud

Note de synthèse

Contexte

Absent plus souvent et plus longtemps, tel est le constat sur le risque arrêt de travail ces dernières années. Selon l'économiste de la santé Frédéric Biard les coûts des indemnités journalières "croient en moyenne de 4% chaque année, soit 500 millions d'euros de plus ".
Plusieurs facteurs peuvent expliquer cette hausse, on peut citer entre autres le vieillissement de la population et l'allongement de l'espérance de vie car l'aide aux personnes agées constitue un facteur d'absentéisme.

Enjeu et problématique

Plusieurs années après l'entrée en vigueur de la directive Solvabilité II, les acteurs du secteur assurantiel ont des progrès à faire en matière de contrôle de la qualité des données utilisées pour calculer les principaux indicateurs prudentiels, comme les provisions techniques, le capital de solvabilité requis (SCR).Telle est la conclusion d'une étude menée en 2019 par l'ACPR [1] auprès de 193 organismes représentant 84 % du chiffre d'affaires du marché de l'assurance et de la réassurance en France.

Les exigences en matière de qualité des données sont mentionnées à l'article 82 de la directive solvabilité 2 et sont rappelées à de nombreuses reprises dans les textes de différents niveaux. L'une des raisons principales pour laquelle la qualité des données est explicitement citée dans le cadre prudentiel et donc associée à une forte vigilence du superviseur est que la donnée constitue la matière première de toute activité d'assurance.

Ce mémoire s'inscrit dans un processus d'amélioration de la qualité des données utilisées lors de l'inventaire durant lequel sont caculées les provisions mathématiques pour chaque individu déclaré en arrêt de travail. Dans un contexte marqué par la volonté de baisse des coûts, la détection de façon automatique des anomalies présentes dans les données qui servent à calculer les provisions est indispensable. En effet des données fiables permettent de produire des résultats plus robustes et dont l'interprétation peut s'avérer plus facile.

Ces anomalies peuvent avoir plusieurs formes parmi lesquelles on peut citer : les fraudes, les doublons, les dossiers provisionnés alors qu'ils sont fermés en prestation, les dossiers topés comme entrant alors que ceux-ci sont déjà présent dans le portefeuille.
Il est d'usage d'effectuer de façon périodique des tâches de revues puis de corrections de ces anomalies, mais ce travail répétitif peut s'avérer coûteux en temps et en ressources.

Dans ce mémoire, il a été proposé une alternative aux méthodes usuelles en utilisant une technique d'apprentissage non-supervisé afin de mettre en évidence les données aberrantes.

1. Autorité de contrôle prudentiel et de résolution

L'approche non supervisée vise à développer un modèle d'apprentissage statistique permettant de regrouper les données en différentes classes homogènes non connues a priori via l'analyse de la structure des données. Elle permet d'identifier donc les données à priori atypiques et sa mise en œuvre est immédiate en ce sens où aucun prérequis n'est nécessaire.
Compte tenu des données et des enjeux le choix s'est porté sur l'algorithme Isolation Forest. Cet algorithme a permis de mettre en évidence les observations présentant un caractère atypique. L'avantage principale de cette approche est la facilité de sa mise en place. Cependant, l'absence de variable cible est son talon d'Achille car elle implique une absence de feedback sur la performance des modèles implémentés. La problématique soulevée est donc de justifier de la pertinence des clusters réalisés.

Des techniques détournées existent pour tester la pertinence des regroupements effectués à savoir :

- évaluation visuelle afin d'observer la répartition des classes dans un espace en 2 dimensions,
- construction d'un modèle d'apprentissage supervisé par-dessus les prédictions réalisées afin de vérifier que les classes sont bien distinctes,
- calcul d'un indice de silhouette.

Application à la détection de fraudes

Les méthodes présentées précédemment ne sont cependant pas robustes, une analyse plus poussée est nécessaire afin de savoir ce qui se cache derrière les anomalies identifiées. L'idée a donc été de tester la robustesse des regroupements effectués via l'Isolation Forest sur une problématique connexe à savoir la détection de fraudes sur les cartes de crédits.

Une seconde base contenant des données étiquetées a donc été étudiée, les résultats obtenus par l'algorithme d'Isolation Forest sont consignés dans le tableau suivant :

	Prédictions	
Classes réelles	0	1
0	283 889	426
1	321	171

Table 1 – Matrice de confusion de l'Isolation Forest sur la détection de fraudes

La classe 0 représente la classe des observations dites "non-frauduleuses" , tandis que la classe 1 représente celle des observations étiquettées comme "frauduleuses".
Les résultats sont assez satisfaisant, en effet sans le moindre paramétrage préalable l'algorithme arrive à détecter près de 35% des observations qui se sont révélées être des fraudes.

Le taux de Faux Positifs est à relativiser pour deux raisons :

— toute fraude peut être considérée comme anomalie mais pas le sens inverse,
— les observations qualifiées comme non-frauduleuses sont en réalité celles dont on a pas réussi à démontrer leur caractère frauduleux,
— le nombre de faux positifs est fortement lié au choix du seuil choisi pour classer les transactions en fonction de leur score d'anomalie.

Ces résultats permettent de justifier la pertinence des regroupements effectués dans le cadre non-supervisé de la détection d'anomalies.

Etant en présence de données étiquetées et dans le soucis d'élargir le cadre de l'étude trois algorithmes d'apprentissage supervisés en plus d'un Auto-Encodeur ont été étudiés il s'agit du Random Forest, XGBOOST et STACKING.

Les résultats de ces algorithmes sur la base de Test sont présentés dans la figure suivante :

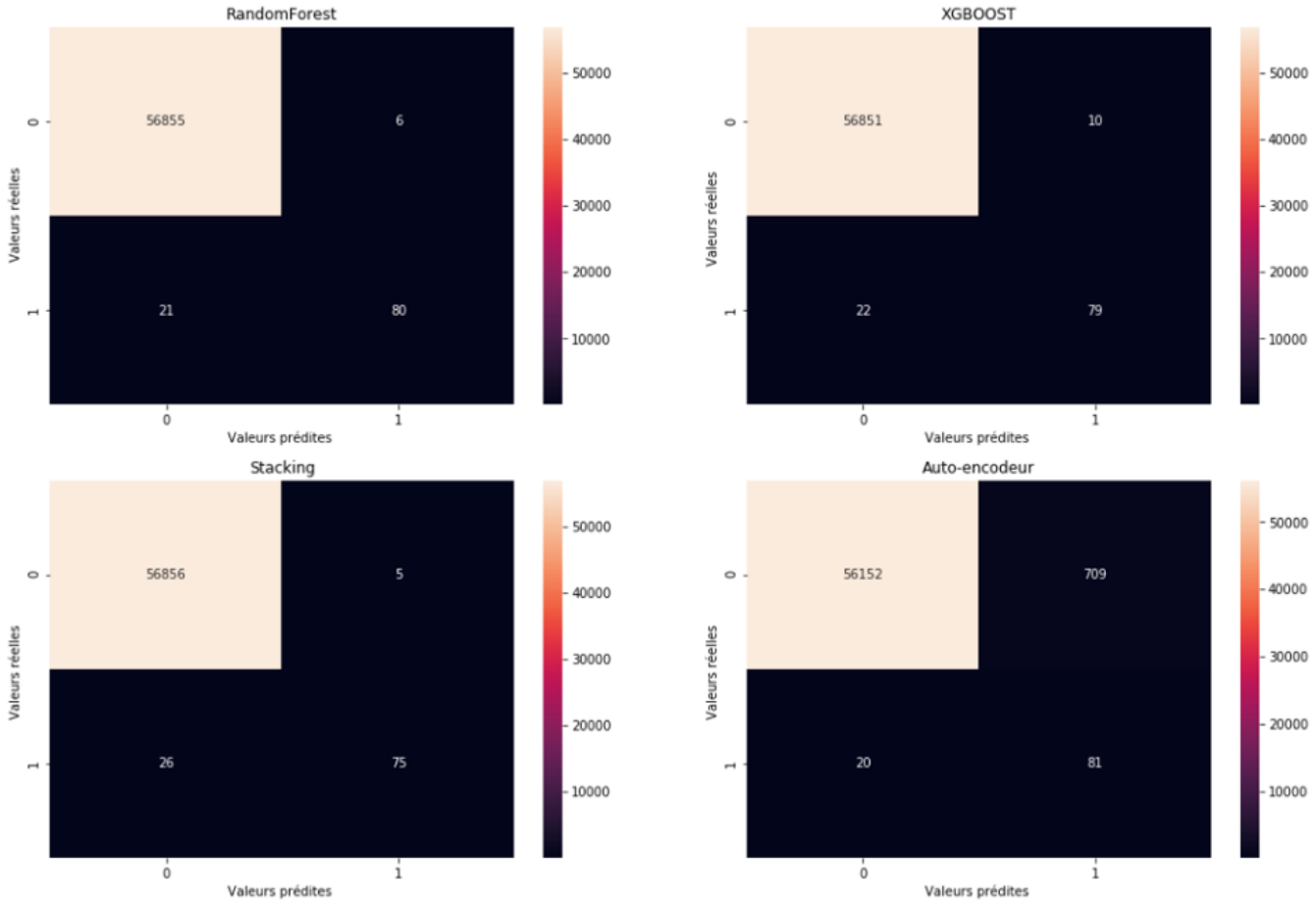
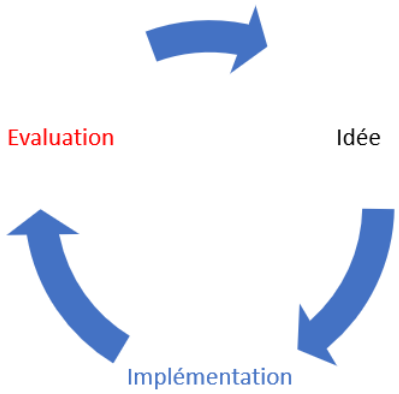


FIGURE 1 – Matrice de confusion des 4 modèles

En observant de plus près les transactions frauduleuses non détectées par les algorithmes d'apprentissage supervisé, il a été osbservé que certaines étaient détectées par l'algorithme d'apprentissage non-supervisé (Isolation Forest).
Explorer une approche mixte combinant ces méthodes d'apprentissage pourrait s'avérer efficace.

La boucle itérative ci-dessous résume bien la démarche entreprise pour combiner ces deux méthodes d'apprentissage :



Chaque idée est mise en place puis évaluée afin d'en tester la pertinence. En procédant de la sorte, le modèle finalement retenu correspond à l'implémentation de l'algorithme Random Forest sur la base rééquilibrée (via l'algorithme SMOTE) et enrichie de la variable liée au score d'anomalie de l'Isolation Forest, ce modèle arrive à identifier 85% des transactions frauduleuses.
Cette approche mixte a permis d'améliorer considérablement les résultats, en effet le score F1 est de 88% soit une hausse significative de 2% en comparaison à l'implémentation du Random Forest sur la base de départ.

Conclusion

Dans le cadre de l'amélioration de la qualité des données, cette étude a permis de démontrer la pertinence des méthodes d'apprentissage non supervisé en particulier de l'Isolation Forest pour la détection d'anomalies et par la même occasion le lien étroit qui existe entre fraudes et anomalies.
Un autre enseignement de cette étude serait d'approfondir les recherches sur la possibilité d'une approche mixte entre l'apprentissage supervisé et l'apprentissage non-supervisé pour modéliser les comportements frauduleux.
En effet, si les algorithmes d'apprentissage supervisé semblent bien adaptés pour détecter les comportements frauduleux déjà observés, un type de fraude jamais détecté ne pourra pas être identifié par ses modèles. L'approche non supervisée pourrait être une alternative efficace pour détecter les nouveaux comportements frauduleux.

Toutefois, une étude de cette envergure nécessite un travail en partenariat avec des experts en charge des sinistres qui par le biais de critères déterministes peuvent mettre en avant certains schémas de fraudes et affiner le cadre de l'étude.
Il pourrait être intéressant de compléter ces résultats en utilisant d'autres méthodes d'apprentissage non-supervisé telles que le LOF [2] ou le DBSCAN [3] et rajouter comme variables explicatives les risques psychosociaux qui constituent un facteur déterminant de l'absentéisme.

2. Local Outlier Factor
3. Density-Bases Spatial Clustering of Applications with Noise

Executive summary

Context

Absent more often and for longer periods, this is the observation on the risk of sick leave in recent years.According to the health economist Frédéric Biard the daily Expenditure due to sick leave "increase by an average of 4% each year, i.e. 500 million euros more".
Several factors can explain this increase, such as the ageing of the population and the increase in life expectancy because assistance to the elderly is a factor in absenteeism

Issues and problems

Several years after the entry into force of the Solvency II directive, players in the insurance sector have progress to make in terms of controlling the quality of the data used to calculate the main prudential indicators, such as technical provisions and the solvency capital requirement (SCR).This is the conclusion of a study carried out in 2019 by the ACPR (French Prudential Supervision and Resolution Authority) on 193 organisations representing 84% of the turnover of the insurance and reinsurance market in France.

Data quality requirements are mentioned in Article 82 of the Solvency 2 Directive and are recalled on numerous occasions in texts at different levels. One of the main reasons why data quality is explicitly mentioned in the supervisory framework and therefore associated with strong supervisory vigilance is that data is the raw material of any insurance activity.

This report is part of a process to improve the quality of the data used during the inventory during which the mathematical provisions for each individual declared to be sick leave are calculated.In a context marked by the desire to reduce costs, the automatic detection of anomalies in the data used to calculate reserving is essential.
Indeed, reliable data means more robust results that can be more easily interpreted.

These anomalies can take several forms, including : fraud, duplication, claims provisioned when they are closed in service, claims marked as incoming when they are already present in the portfolio.
It is common practice to periodically review and then correct these anomalies, but this repetitive work can be time and resource consuming.

In this report, an alternative to the usual methods has been proposed using unsupervised learning techniques to highlight outliers.
The unsupervised approach aims to develop a statistical learning model that allows the data to be grouped into different homogeneous classes that are not known a priori via the analysis of the data structure. It therefore makes it possible to identify a priori atypical data and its implementation is immediate in the sense that no prerequisites are necessary.

Given the data and the issues at stake, the Isolation Forest algorithm was chosen.This algorithm made it possible to highlight the observations presenting an atypical character. The main advantage of this approach is the ease with which it can be implemented. However, the absence of a target variable is its Achilles' heel as it implies an absence of feedback on the performance of the implemented models. The problem raised is therefore to justify the relevance of the clusters implemented.

Backdoor techniques exist to test the relevance of the groupings made, namely :

- Visual assessment to observe the distribution of classes in a 2 dimensional space.
- Construction of a supervised learning model on top of the predictions made in order to verify that the classes are indeed distinct.

Application to fraud detection

However, the methods presented above are not robust, and further analysis is needed to find out what lies behind the anomalies identified.The idea was therefore to test the robustness of the groupings carried out by the Isolation Forest on a related problem, namely the detection of credit card fraud.

A second database containing labelled data was thus studied, the results obtained by the Isolation Forest algorithm are recorded in the following table :

	Prédictions	
Classes réelles	0	1
0	283 889	426
1	321	171

Table 2 – Isolation Forest's confusion matrix on fraud detection

Class 0 represents the class of observations labelled as "non-fraudulent", while class 1 represents the class of observations labelled as "fraudulent".
The results are quite satisfactory, indeed without any prior parameterization the algorithm manages to detect nearly 35% of the observations that turned out to be fraudulent.

The rate of False Positives should be put into perspective for two reasons :

— Each fraud can be considered as an anomaly but not the other way around
— Observations classified as non-fraudulent are in fact those that have not been shown to be fraudulent

These results justify the relevance of the groupings made in the unsupervised framework of anomaly detection.

In the presence of labelled data and in order to widen the scope of the study, 3 supervised learning algorithms in addition to an Auto-Encoder were studied : Random Forest, XGBOOST and STACKING.

The results of these algorithms on the Test base are shown in the following figure :

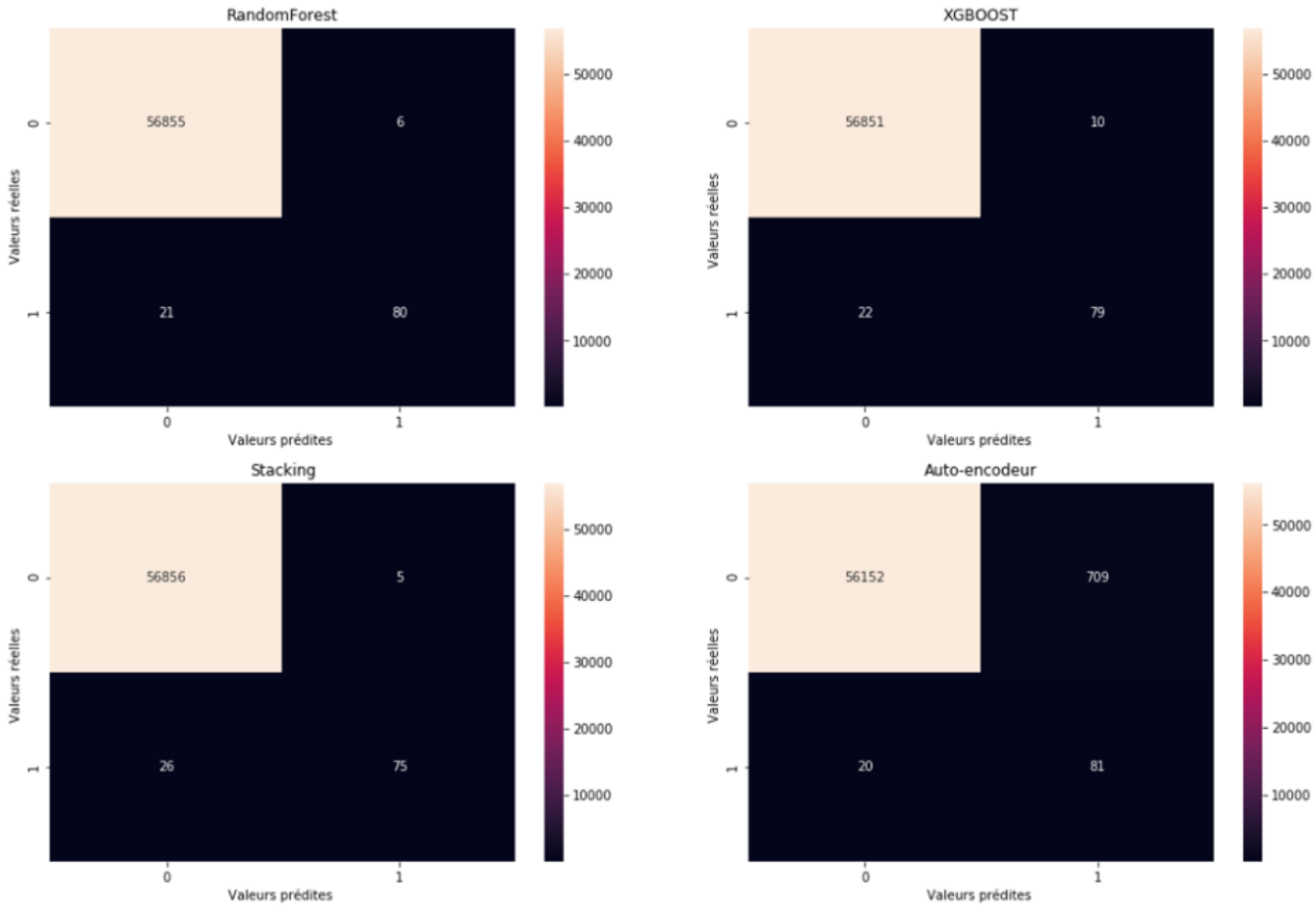
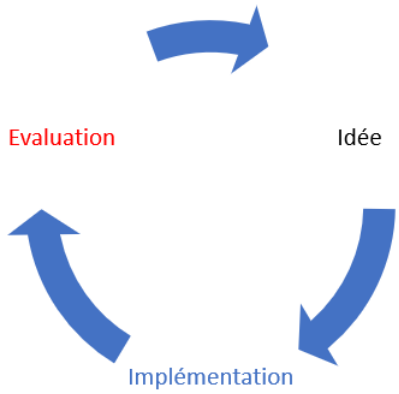


FIGURE 2 – Confusion matrix of the 4 models

Looking more closely at the fraudulent transactions not detected by the supervised learning algorithms, it was observed that some were detected by the unsupervised learning algorithm (Isolation Forest). Exploring a mixed approach combining these learning methods could be effective.

The iterative loop below summarises the approach taken to combine these two learning methods :



Each idea is implemented and then evaluated to test its relevance. By proceeding in this way, the model finally retained corresponds to the implementation of the Random Forest algorithm on the rebalanced base (through the SMOTE algorithm) and enriched with the feature of the anomaly score of the Isolation Forest, this model manages to identify 85% of fraudulent transactions.
This mixed approach allowed for a considerable improvement in the results, as the F1 score is 88%, a significant increase of 2 percentage points compared to the implementation of the Random Forest on the initial database.

Conclusion

In the context of data quality improvement, this study has demonstrated the relevance of unsupervised learning methods, in particular Isolation Forest, for anomaly detection at the same time the close link between fraud and anomalies.
Another lesson from this study was be to further investigate the possibility of a mixed approach between supervised and unsupervised learning to model fraudulent behaviour.
Indeed, if supervised learning algorithms seems to be well adapted to detect fraudulent behaviours already observed, a type of fraud never detected could not be identified by these models. The unsupervised approach could be an efficient alternative to detect new fraudulent behaviour.

However, a study of this magnitude requires working in partnership with claims adjusters who, through deterministic criteria, can highlight certain patterns of fraud and refine the framework of the study.
It might be interesting to complement these results by using other non-supervised learning methods and adding psychosocial risks as an explanatory variable, which are determining factor in absenteeism.

Remerciements

Tout d'abord, je tiens à remercier Jérémy Boulay et Sébastien Vitry de m'avoir acceuilli au sein du service et de m'avoir offert la possibilité de travailler sur ce sujet.

Je présente également mes plus sincères remerciements à tout le service inventaire santé prévoyance dans son intégralité pour son accueil et pour l'aide reçue qui fut précieuse.

J'adresse également ma gratitude à toutes les personnes qui m'ont apporté leur aide tout au long de ce mémoire. En particulier Mohamed Soumahoro, Bakayoko Abdel-Aziz, Vitalii Chugui et Amina Nimour.

Je remercie mon tuteur, monsieur Franck Vermet directeur de l'EURIA pour sa relecture et ses conseils avisés. Mes remerciements vont également à Patricia Gruet pour son soutien.

La rédaction du mémoire d'actuariat étant un exercice complexe, dans lequel l'appui des proches est indispensable. Aussi, je remercie ma famille et mon proche entourage, avec une pensée particulière à ma compagne et ma mère pour leurs encouragements et leur soutien.

Table des matières

I Introduction **3**

1 Prévoyance en France 4

1.1 Définition 4

1.2 Les garanties prévoyance des régimes complémentaires 5

2 Focus sur l'arrêt de travail 6

2.1 Description des états 7

2.2 Provisions mathématiques 8

2.3 Calcul des provisions 8

3 Sinistralité en arrêt de travail 10

4 Un mot sur la qualité des données 11

5 Démarche de l'étude 13

II Détection d'anomalies via l'apprentissage non supervisé **14**

1 Construction de la base **15**

1.1 Etapes préliminaires 15

1.2 Feature Engineering 16

1.2.1 Définition 16

1.2.2 Nouvelles variables 16

1.3 Rajout de variables externes 19

1.3.1 Constat 19

1.3.2 Variables externes 19

2 Analyse exploratoire de la base **23**

2.1 Analyse de la forme 23

2.1.1 Typologie des variables 23

2.1.2 Valeurs manquantes 23

2.2 Analyse du fond 24

2.2.1 Focus sur l'âge et le sexe 24

2.2.2 Lien entre durée en arrêt, le motif et le sexe 26

2.2.3 Focus sur la zone 27

2.2.4 Lien entre invalidité et l'âge 28

2.2.5 Saisonnalité 29

3 Modélisation **31**

3.1 Rappels théoriques du modèle utilisé 32

3.2 Application de l'algorithme 36

3.2.1 Preprocessing 36

3.2.2 Modélisation 36

3.3 Présentation des résultats 38

3.4 Procédure d'évaluation 39

3.5 Conclusion 42

III Application à la détection de fraudes **43**

4 Etude de la base à disposition **46**

4.1 Présentation de la base 46

4.2 Statistiques sur la base à disposition 46

4.2.1 Relation entre les variables explicatives et la variable cible 47

4.2.2 Relation entre variables explicatives 49

5 Modélisation **51**

5.1 Apprentissage supervisé 51

5.2 Mesure de la qualité des modèles 52

5.2.1 Dilemme biais-variance 52

5.2.2 Notions de sur-apprentissage 54
5.2.3 Validation croisée 56
5.2.4 Mesure de performance 57
5.3 Choix des algorithmes 60
5.3.1 Comment choisir un bon modèle ? 60
5.3.2 Présentation théorique des modèles 61
5.4 Présentation des résultats 74
5.4.1 Apprentissage non supervisé 74
5.4.2 Apprentissage supervisé 76
5.5 Amélioration des performances : mixte entre l'apprentissage supervisé et la vision non-supervisée 78
5.5.1 Choix du modèle à améliorer et justification des approches considérées 79
5.5.2 Méthode 1 : Rééquilibrage des données et utilisation du score d'anomalie de l'Isolation Forest 79
5.5.3 Méthode 2 : Prédiction via l'espace latent de l'Auto-Encodeur 82
5.6 Synthèse des résultats 83
5.7 Conclusion 85

Table des figures

1 Matrice de confusion des 4 modèles v

2 Confusion matrix of the 4 models ix

3 Structuration de la protection sociale Française 4

4 Cycle des états d'un assuré en arrêt de travail 7

5 Le cycle d'évolution en incapacité et en invalidité 8

6 Evolution des dépenses d'indemnités journalières maladie, du nombre de journées indemnisées et de la masse salariale plafonnée à 1,8 SMIC (2010-2017, base 100 en 2010) 10

7 Nombre d'arrêts et dépenses d'indemnisation en fonction de la durée des arrêts maladie (2011 et 2017) 11

8 Solvabilité 2 et la qualité des données 12

1.1 Illustration de la troncature et de la censure 18

2.1 Représentation de l'architecture de la base en fonction des valeurs manquantes 24

2.2 Histogramme de la variable âge en fonction du sexe 25

2.3 Distribution des arrêts de travail de sexe Féminin entre 18 et 39ans 25

2.4 Durée de l'arrêt en fonction du sexe 26

2.5 Nombre d'arrêts de travail déclarés en fonction du département 28

2.6 Tableau croisé entre la variable Top Invalidité et la variable âge 28

2.7 Sinistralité par an en fonction des 12 mois de l'année 29

2.8 Les motifs d'arrêts par mois 30

3.1 Segmentation en deux dimensions 33

3.2 Exemple d'une isolation tree, cas bidimensionnel 34

3.3 Evolution du score d'anomalie en fonction de E(h(x)), source[12] 36

3.4 Histogramme des scores d'anomalie 37

3.5 Présentation des résultats des clusters pour différentes variables 38
3.6 Projection de 3 points dans un espace 2d 40
3.7 Réduction de l'espace en une dimension 40
3.8 Représentation en deux dimensions des clusters réalisés via l'algorithme d'Isolation Forest 41

4.1 Distribution des variables en fonction du type de classe 47
4.2 Matrice de corrélation des variables 50

5.1 Illustration du compromis biais-variance 53
5.2 Illustration d'un modèle en sur-apprentissage 54
5.3 Lien entre complexité et erreur de prédiction, source [14] 55
5.4 Découpage du jeu de données 56
5.5 Cross validation avec un nombre de split=5 57
5.6 Matrice de confusion 58
5.7 Courbe ROC de différents modèles 59
5.8 Carte des algorithmes de scikit-learn, source [18] 60
5.9 Algorithme de bagging, source [1] 62
5.10 Présentation de l'algorithme Random Forest, source [1] 63
5.11 Exemple d'un schema de stacking 68
5.12 Illustration de l'architecture d'un Autoencodeur, source [19] 69
5.13 Comparaison d'un ACP et d'un auto encodeur en 2 dimensions 71
5.14 Illustration d'un auto-encodeur variationnel sous l'hypothèse d'une gaussienne multivariée, source [19] 71
5.15 Schéma d'un auto-encodeur 73
5.16 Histogramme des scores d'anomalies 75
5.17 Matrice de confusion des 4 modèles 77
5.18 Description du suréchantillonage via la méthode SMOTE, source [8] 80
5.19 Vision en 2D de la segmentation via l'algorithme K-Prototype 91
5.20 Illustration de la notion de point central, bordure et bruit 93
5.21 Décomposition de l'évolution du montant d'indemnités journalières par la CNAM, source[15] 98
5.22 Fonction ReLU 99

5.23 Fonction tanh 99
5.24 Fonction Sigmoïde 100

Introduction générale

Dans un monde du Big Data, les problèmes liés à la qualité des données ne font que croître de façon exponentielle.

La qualité des données est le processus de maintien de la cohérence et de la fiabilité des valeurs détenues dans le système informatique d'une entreprise. Une information qualitative répond à plusieurs critères comme la validité, la précision, la cohérence, la complétude et la mise à jour.
A l'inverse, la non-qualité des données représente l'absence de qualification de celles-ci. Elle regroupe l'indisponibilité, la non-conformité, la non-actualisation, la non-complétion et l'inexactitude des valeurs. Le plus souvent une information non qualitative se manifeste par :

— des données inconnues et inexactes,
— des données non ou mal définies,
— des données redondantes et incohérentes.

Plusieurs études sur la data quality convergent sur des conclusions alarmantes :

— selon une étude de l'ACPR menée en 2019, les acteurs du secteur assurantiel ont des progrès à faire en matière de contrôle de la qualité des données utilisées pour calculer les principaux indicateurs prudentiels, comme les provisions techniques, le capital de solvabilité requis (SCR),
— une étude du MIT révèle que : la non-qualité occasionne une perte d'argent estimée entre 15% et 25% du chiffre d'affaire total d'une entreprise,
— une étude de CrowdFlower datant de 2016 estime que 80% des spécialistes de la data science déclarent perdre leur temps en traitant des données de mauvaise qualité,
— selon une étude du laboratoire de recherche d'AT&T, la pauvre qualité des données coûte aux entreprises américaines 600 milliards de dollars par an. Cela n'englobe pas seulement le préjudice lié à des erreurs graves découlant d'une mauvaise qualité des données, mais aussi la quantité de temps et les laborieux efforts humains nécessaires pour corriger ces données.

Pourtant, les données sont au coeur de toutes les études actuarielles. Tout écart matériel entre les données réelles et celles utilisées pour des calculs divers peut aboutir à des résultats très éloignés de la réalité et les analyses qui en resulteront le seront également.

On peut donc en conclure que la qualité des données est un élément essentiel du pilotage de l'entreprise, les exigences en la matière sont d'ailleurs mentionnées à l'article 82 de la directive Solvabilité II et

sont rappelées à de nombreuses reprises dans des textes de différents niveaux.

Dans ce mémoire, nous nous sommes intéressés aux anomalies présentes dans la base servant au calcul des provisions mathématiques en arrêt de travail. La détection de ces anomalies a été réalisée dans un cadre non-supervisé via l'algorithme d'Isolation Forest. Il est important de noter que l'application de cet algorithme peut être généralisée à d'autres types de bases de données.
Dans un second temps, dans le but de verifier la pertinence de la segmentation effectuée, l'Isolation Forest a été implémenté dans un processus de detection de fraudes. En l'absence de données sur la fraude en arrêt de travail, cette détection a eté réalisée sur une seconde base relative à la problématique connexe de fraudes sur les cartes de crédits.

Pour compléter l'étude, une approche mixte alliant la vision supervisée adaptée pour les fraudes déjà observées et une vision non supervisée dont l'objectif serait de mettre en lumière les nouveaux types de fraudes a été mise en place. Cette approche combinée a été développée dans le but de pouvoir prendre en compte la structure évolutive de la fraude dans le temps.

Première partie

Introduction

1 Prévoyance en France

1.1 Définition

Selon la loi n° 89-1009 du 31 décembre 1989, dite loi EVIN, la prévoyance regroupe « les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque chômage ». Les prestations de prévoyance sont organisées selon plusieurs niveaux résumés dans le schéma suivant :

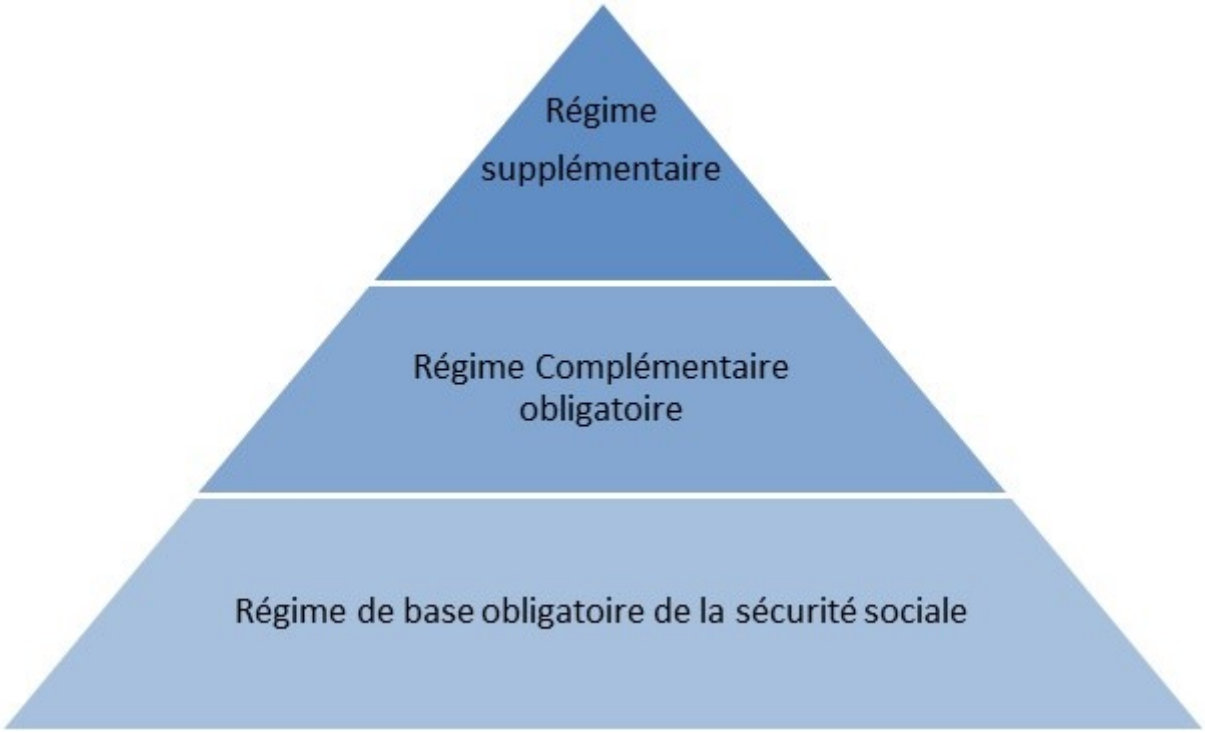


FIGURE 3 – Structuration de la protection sociale Française

Regime de Base

Il couvre les risques ou événements ayant pour conséquence une perte de revenus : maladie, maternité, accident, perte d'emploi ou décès. Une prestation compensatoire est versée, après un délai de carence de trois jours, pour palier la diminution de salaire due à l'impossibilité de travailler. Elle est égale à 50% du salaire journalier de base [4] et dans la limite de 1,8 fois le SMIC. Ce régime est financé par les cotisations sociales obligatoires prélevées sur la paye de chaque salarié. Des régimes complémentaires s'ajoutent pour compléter ce régime obligatoire.

4. moyenne des 3 salaires bruts des 3 derniers mois précédent l'arrêt de travail, ou des 12 derniers mois d'activité en cas d'activité saisonnière ou discontinue.

Régime Complémentaire obligatoire

En matière de prévoyance, les entreprises doivent répondre à des obligations légales et conventionnelles :

— La loi du 19 janvier 1978 dite loi de « mensualisation » oblige les employeurs à assurer (sous certaines conditions et avec un délai de carence de 7 jours), un certain niveau de salaire en cas d'arrêt de travail pour maladie ou accident aux salariés ayant au moins un an d'ancienneté : 90% du salaire brut pendant les 30 premiers jours et 66,66% du salaire brut pendant les 30 jours suivants (plus 10 jours supplémentaires par tranche d'ancienneté de 5 ans),
— La convention collective nationale pour les cadres de mars 1947 impose aux employeurs de souscrire une garantie décès pour les cadres,
— Les employeurs qui adhèrent à une convention ou accord selon leur type d'activité ou leur branche d'activité sont tenus d'appliquer les garanties obligatoires en prévoyance si celles-ci sont mentionnées dans l'accord.

Régime supplémentaire

Les contrats peuvent être souscrits dans le cadre de l'entreprise et répartis entre les mutuelles, les sociétés d'assurance et les institutions de prévoyance. A titre individuel, il est possible de souscrire un contrat repondant au manque à gagner qu'engendrerait une incapacité ou un décès. Ces produits s'adressent principalement aux Travailleurs Non-Salariés (TNS) car leur régime obligatoire ne les couvre que très faiblement en comparaison avec les salariés pouvant être couverts par le régime de prévoyance collective de leur entreprise.

1.2 Les garanties prévoyance des régimes complémentaires

1.2.1 Les garanties prévoyance en cas d'arrêt de travail

Des indemnités journalières sont versées en complément des paiements effectués par la sécurité sociale, la garantie permet donc le maintien total ou partiel du salaire en cas d'arrêt de travail pour maladie ou accident. L'arrêt de travail se ventile en deux types d'état : l'incapacité et l'invalidité.

Etat incapacité

L'incapacité est l'état d'une personne qui, par suite de maladie ou d'accident, se trouve dans l'impossibilité de continuer son activité professionnelle. De manière générale, l'incapacité de travail est définie à l'article 321-1 du code de la sécurité sociale, par l'incapacité physique, constatée par un médecin, de continuer ou de reprendre son activité professionnelle. Dans ce cas de figure, la garantie prévoit le versement à l'assuré en état d'incapacité d'une indemnité journalière en pourcentage du salaire annuel brut. Le niveau de salaire garanti dépend de la nature du sinistre (maladie, accident du travail). La

durée d'indemnisation maximale est de 36 mois. Afin d'éviter tout effet d'enrichissement, le montant des prestations cumulées (sécurité social et complémentaires) ne peut excéder le salaire annuel perçu par l'assuré avant l'arrêt de travail.

Etat invalidité

Est considéré comme invalide au sens de la sécurité social, toute personne qui à la suite d'un accident ou une maladie, voit sa capacité de travail ou de gain réduite d'au moins deux tiers. Les individus sont alors classés en 3 catégories [5] :

— 1ère catégorie : invalides capables d'exercer une activité rémunérée (pouvant être différente de celle exercée auparavant),
— 2ème catégorie : invalides absolument incapables d'exercer une profession quelconque
— 3ème catégorie : invalides qui, étant absolument incapables d'exercer une profession, sont, en outre, dans l'obligation d'avoir recours à l'assistance d'une tierce personne pour effectuer les gestes simples de la vie quotidienne.

Si l'assuré est classé par la sécurité sociale dans l'une des trois catégories d'invalides, alors il perçoit une rente invalidité basée sur un pourcentage du salaire brut de l'assuré. De plus, le montant de la rente versée dépend du niveau d'incapacité de la victime. La prestation d'invalidité prend fin dès lors que l'assuré part à la retraite ou décède. Dans le cas d'une personne en incapacité l'état de sortie peut être lié à plusieurs facteurs : la reprise du travail, son passage en invalidité, son départ à la retraite, son décès.

1.2.2 Garantie décès

La prestation résultant de la garantie décès peut prendre différentes formes. Elle est fonction du salaire brut annuel perçu au cours des 12 derniers mois précédent le décès, mais dépend également des caractéristiques familiales de l'assuré (nombre d'enfants, conjoint...)

2 Focus sur l'arrêt de travail

Maintenant que les garanties liées à l'arrêt de travail ont été présentées il est nécessaire de comprendre les risques auxquels l'assureur est exposé dans ce type de garanties.

5. Ces 3 catégories sont définies à l'article L.341-4 du code de la Sécurité Sociale.

2.1 Description des états

L'assureur doit être capable de faire face à ses engagements futurs. Pour se faire, il doit appréhender au mieux les différents états d'un assuré en arrêt. Le graphique ci-dessous représente le cycle d'évolution d'un état à un autre.

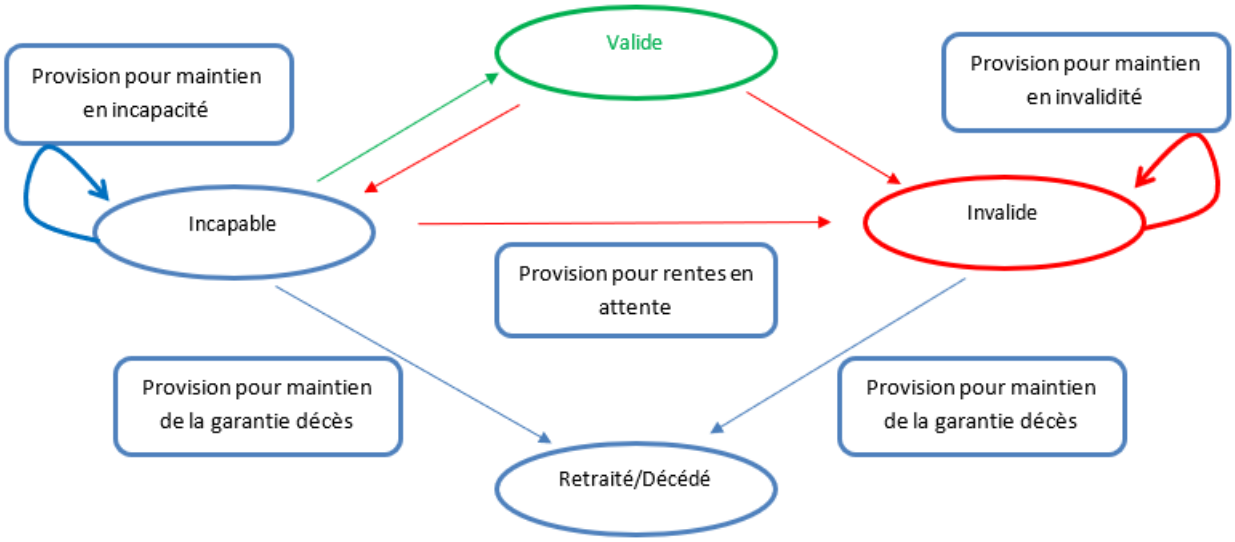


FIGURE 4 – Cycle des états d'un assuré en arrêt de travail

Plusieurs possibilités peuvent se présenter lorsqu'un assuré est en arrêt pour incapacité. Il peut ensuite passer dans les états suivants :

- reprise du travail,
- passage en invalidité,
- départ à la retraite,
- décès.

Une personne en invalidité peut sortir de son état par :

- un départ à la retraite,
- son décès.

Un individu en état d'invalidité ne peut voir son état s'améliorer (impossibilité de repasser en incapacité voir devenir redevenir valide), l'invalidité est considérée comme irréversible .

2.2 Provisions mathématiques

Elles représentent la valeur actuelle probable des prestations futures de l'assureur au titre des sinistres connus à la date d'inventaire. Elles s'inscrivent en 4 catégories :

- La provision incapacité en cours : elle doit permettre de couvrir l'engagement de l'assureur lié au maintien de l'assuré en incapacité,
- La provision invalidité en attente : elle correspond aux rentes d'invalidité censées couvrir l'engagement de l'assureur en cas de passage d'un incapable en invalidité ,
- La provision invalidité en cours : son rôle est de couvrir l'engagement de l'assureur lié au maintien de l'assuré en invalidité,
- La provision de maintien de la garantie décès : elle permet de couvrir le versement de prestations en cas de décès de l'assuré en arrêt de travail.

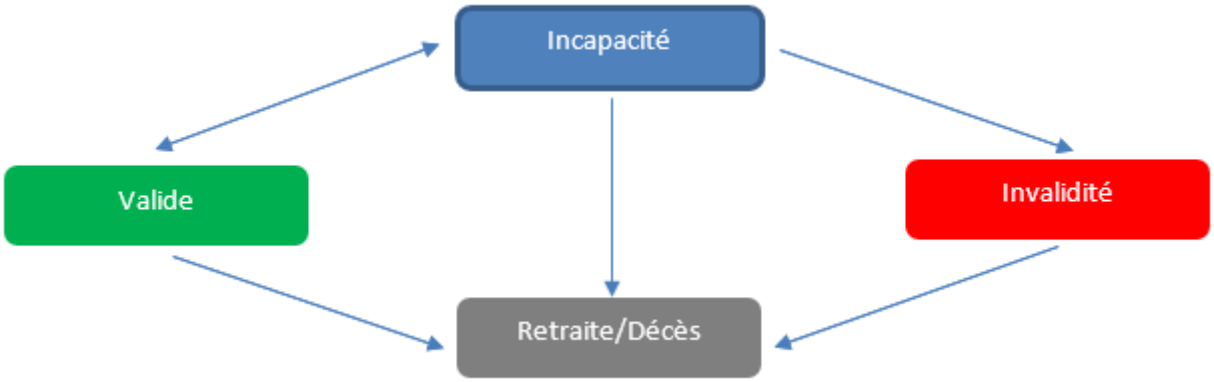


FIGURE 5 – Le cycle d'évolution en incapacité et en invalidité

A partir du schéma ci-dessus nous pouvons en déduire que la Provision mathématique totale d'incapacité correspond à la somme des provisions de maintien en incapacité, d'invalidité en attente et de maintien de la garantie décès.

2.3 Calcul des provisions

L'objectif ici est de présenter les formules de calculs tête par tête pour une rente annuelle de 1€ des trois provisions suivantes :

- la provision pour maintien en incapacité,
- la provision pour invalidité en attente,
- la provision pour maintien en invalidité.

0.2.3.1 La provision pour maintien en incapacité

Cette provision doit permettre de couvrir l'engagement de l'assureur lié au maintien de l'assuré en incapacité. La formule de calcul de cette provision est la suivante :

$$PM^{INC}(x,a) = \sum_{k=a+1}^{36} \frac{l^{INC}(x,k)}{l^{INC}(x,a)} \times \frac{1}{(1+i_m)^{(k-a)}}$$

avec :
x : âge d'entrée dans l'état d'incapacité.
a : nombre de mois écoulés depuis l'entrée dans l'état d'incapacité.
$l^{INC}(x,k)$: loi de maintien en incapacité pour un individu entré en incapacité à l'âge x et qui est en incapacité depuis k mois.
i_m : taux d'actualisation mensuel qui est obtenu à partir du taux d'actualisation annuel de la manière suivante :

$$(1+i) = (1+i_m)^{12} \iff i_m = (1+i)^{\frac{1}{12}} - 1$$

i : taux annuel.

0.2.3.2 La provision pour invalidité en attente

Cette provision doit permettre de couvrir l'engagement de l'assureur lié au risque de passage de l'assuré en invalidité, sachant qu'il est actuellement en état d'incapacité. La formule de calcul de cette provision est la suivante :

$$PM^{INV}_{att}(x,a) = \sum_{k=a+1}^{36} \frac{1}{(1+i_m)^{k-a}} \times \frac{s(x,k-1)}{l^{INC}(x,a)} \times PM^{INV}(x+\frac{k}{12},0)$$

avec :
x : âge d'entrée dans l'état d'incapacité.
a : nombre de mois écoulés depuis l'entrée dans l'état d'incapacité variant de 1 à 36 mois.
$l^{INC}(x,k)$: loi de maintien en incapacité pour un individu entré en incapacité à l'âge x et qui est en incapacité depuis k mois.
$s(x,k)$: loi de passage en invalidité pour un individu entré en incapacité à l'âge x et qui est en incapacité depuis k mois.
$PM^{INV}(x+\frac{k}{12},0)$: provision d'invalidité pour un individu de 0 mois d'ancienneté en invalidité et d'âge x+ k/12.
i_m : taux d'actualisation mensuel obtenu à partir du taux d'actualisation annuel.

0.2.3.3 la provision pour maintien en invalidité

Elle doit permettre de couvrir l'engagement de l'assureur lié au maintien de l'assuré en invalidité. Elle se calcule de la façon suivante pour une garantie couvrant l'invalidité jusqu'à 62 ans.

$$PM^{INV}(x,a) = \sum_{k=a+1}^{62-x} \frac{l^{INV}(x,k)}{l^{INV}(x,a)} \times \frac{1}{(1+i)^{k-a}}$$

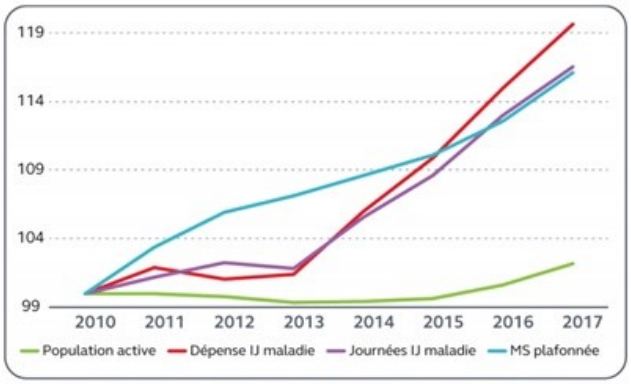
avec,
x : âge d'entrée dans l'état d'invalidité.
a : nombre d'années écoulées depuis l'entrée dans l'état d'invalidité.
$l^{INV}(x,k)$: loi de maintien en invalidité pour un individu entré en invalidité à l'âge et qui est en invalidité depuis k années.
i : le taux d'actualisation annuel pour l'année de projection k.

3 Sinistralité en arrêt de travail

Nous allons nous intéresser dans cette section à la sinistralité observée au niveau national en arrêt de travail.

Les chiffres donnés dans cette section proviennent du rapport de la cour des comptes sur la sécurité sociale paru en Octobre 2019.

Une vive progression des dépenses d'indemnités journalières est observée depuis 2013. Cette hausse de sinistralité est portée par les indemnités journalières pour maladie et par les accidents du travail et maladies professionnelles. Le graphique ci-dessous illustre ce constat.



Source : Cour des comptes, d'après des données de l'Insee, de la DSS et de la CNAM.

FIGURE 6 – Evolution des dépenses d'indemnités journalières maladie, du nombre de journées indemnisées et de la masse salariale plafonnée à 1,8 SMIC (2010-2017, base 100 en 2010)

La masse salariale plafonnée à 1,8 SMIC est en constante évolution depuis 2013. Elle évolue sous l'effet de la hausse de la population active, du niveau des salaires et du SMIC. Cette évolution reste néanmoins plus faible que celle de la hausse des dépenses d'indemnités journalières liées à la maladie. L'écart entre ces deux évolutions met en évidence l'importance d'un autre paramètre influent sur la dépense à savoir le nombre moyen de jours d'arrêt de travail par salarié.

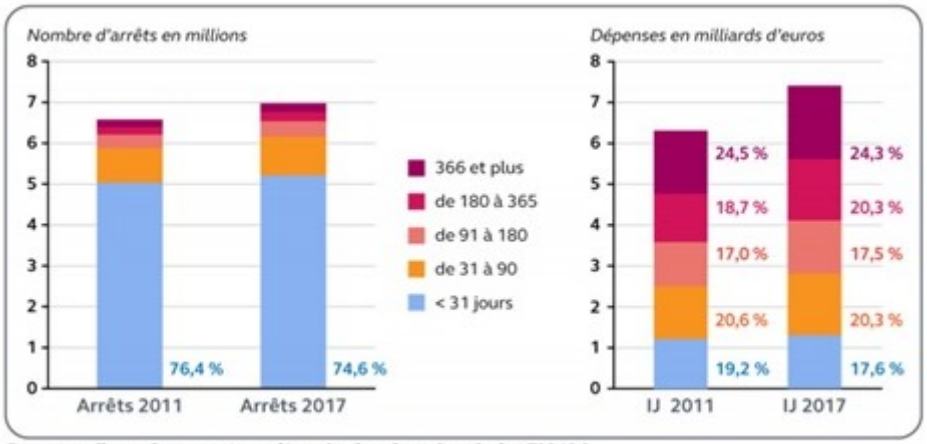


FIGURE 7 – Nombre d'arrêts et dépenses d'indemnisation en fonction de la durée des arrêts maladie (2011 et 2017)

Pour étayer ces observations, la figure ci-dessus montre que la dérive de sinistralité observée est non seulement liée à un effet volume(nombre d'arrêts) mais aussi à un effet d'inflation. En effet, la part des arrêts de plus de 6 mois représentait 5,5% des arrêts et 43,3 % en 2011. En 2017, le pourcentage des arrêts de plus de 6 mois est de 6,2 % ce qui représente 44,6 % des dépenses.

4 Un mot sur la qualité des données

La qualité des données est un élément essentiel du pilotage de l'entreprise, abordée dans les trois piliers réglementaires de Solvabilité 2 ce sujet est donc transversal.
Les exigences en matière de qualité des données sont mentionnées à l'article 82 de la directive solvabilité II et sont rappelées à de nombreuses reprises dans des textes de différents niveaux comme présenté dans la figure suivante :

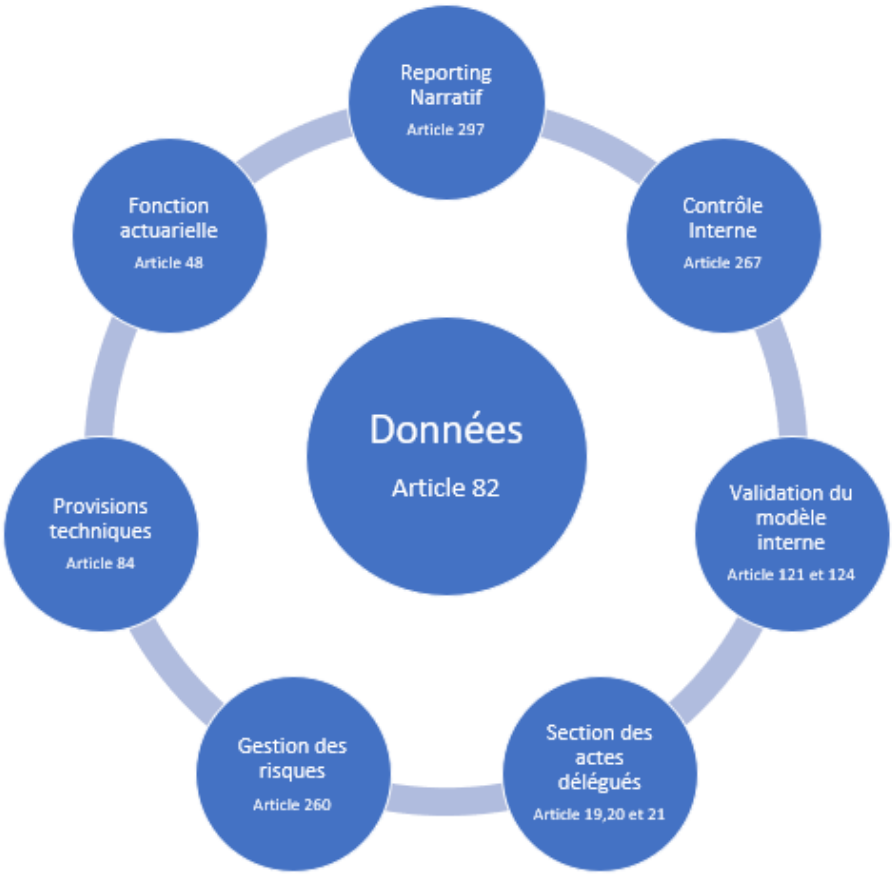


FIGURE 8 – Solvabilité 2 et la qualité des données

Le principal objectif de ce dispositif est de garantir la fiabilité des informations donc de refléter l'image fidèle des engagements des assureurs. Les risques encourus en cas de défaillance du dispositif de la qualité des données sont nombreux à savoir :

- augmentation des coûts opérationnels (immobilisation inutile de capital),
- risque d'initier une mauvaise stratégie et de manquer des opportunités commerciales,
- baisse de la qualité/ pertinence des analyses,
- non-conformité aux éxigences réglementaires (risque de pénalité du régulateur).

Le régulateur pose ses exigences en matière de qualité des données autour de trois principaux critères :

- exactitude : les données ne doivent pas être affectées d'erreurs matérielles, elles doivent être cohérente dans le temps,
- exhaustivité : elle se traduit par une granularité et une profondeur suffisante pour identifier les tendances et l'évolution des riques sous-jacents,
- pertinence : la pertinence des données siginifie qu'elles doivent refléter les risques auxquels est exposé l'assureur au travers de ses engagements.

5 Démarche de l'étude

Compte tenu de l'absence de données labélisées à disposition, la deuxième partie de ce mémoire a été consacrée à la détection d'observations atypiques dans un cadre non-supervisé sur la base utilisée pour calculer les provisions mathématiques en arrêt de travail.
L'approche non supervisée vise à développer un modèle d'apprentissage statistique permettant de regrouper les données en différentes classes homogènes non connues a priori via l'analyse de la structure des données. Elle permet donc de mettre en évidence les observations présentant un caractère atypique.
L'avantage principale de cette approche est la facilité de sa mise en place opérationnelle. Cependant, l'absence de variable cible est le talon d'Achille de l'approche car elle implique une absence de feedback sur la performance des modèles implémentés.
La problématique soulevée est de justifier de la pertinence des différents regroupements effectués. Cette pertinence peut se mesurer au travers de l'information supplémentaire que ces clusters apporteraient sur la base de données.
Par exemple, imaginons le cas où les observations atypiques s'avèrent être des fraudes la segmentation réalisée serait d'un grand intérêt. Cette idée n'est pas anodine en ce sens où toute observation frauduleuse de par sa structure revêt un caractère atypique.
Malheureusement, nous ne disposons pas de base labélisée sur la fraude en arrêt de travail raison pour laquelle une seconde base a été étudiée sur la problématique connexe de détection de fraudes sur les cartes de crédits.
L'algorithme présenté dans la phase d'apprentissage non-supervisé sera implémenté puis testé afin de mesurer son pouvoir de détection de transactions frauduleuses.

Le language de programmation utilisé à partir du chapitre 2 est Python.

Deuxième partie

Détection d'anomalies via l'apprentissage non supervisé

Chapitre 1

Construction de la base

Comme dans tout problème de data-science les données sont le cœur du sujet. L'objectif de cette partie est d'arriver à fournir une base de données fiables et en quantité suffisante afin de d'assurer la robustesse et la précision des résultats de la modélisation.
La base de données utilisée dans le cadre de cette étude est le résultat de jointure de plusieurs bases SAS disponible dans l'entrepôt de données. Pour des soucis d'obtention de données fiables et complètes, les arrêts de travail en délégation ont été retirés de la base de données. Le souhait était également de travailler sur des années complètes avec une profondeur d'historique assez conséquente. L'étude portera donc sur les dossiers arrêts de travail pour les contrats collectifs à adhésion obligatoire gérés en gestion interne sur les exercices de survenance de 2012 à 2019.

1.1 Etapes préliminaires

Il a fallu procéder à un certain nombre de retraitements sur la base brute issue de la jointure pour l'adapter à la problématique de notre étude.

Suppression des doublons

Suppression des lignes qui se répètent via la clé numéro de sécurité sociale/date d'arrêt car elles peuvent constituer un biais dans notre analyse.

Sélection des arrêts de travail survenus entre 18 ans et 65 ans

La population active considérée ici est celle des 18-65 ans.

Suppression des arrêts de travail ayant une franchise supérieure à 365 jours

On conserve les arrêts ayant une franchise maximum de 365 jours , afin de ne pas inclure des assurés non représentatifs du marché et dont les comportements peuvent différer du reste de la population.

Suppression des arrêts $>$ 1095 jours

On ne conserve que les individus en incapacité.

Identification des arrêts ayant pour motif de sortie le passage en invalidité

On récupère la base contenant les informations liées aux rentes invalidité et on réalise un rapprochement pour identifier les assurés ayant pour motif de sortie le passage en invalidité.

Correction des anomalies

Les sinistres dont les dates sont non renseignées, ou ceux dont la date d'arrêt est antérieure à la date d'effet du contrat ont été écartés. Les assurés ayant plusieurs dates de mise en invalidité ce qui biaise l'analyse du comportement individuel des assurés ont aussi été écartés. Les arrêts dont la date de réception des justificatifs est antérieure à la date d'arrêt ont été retraités, par soucis de simplification ils ont été catégorisés en sinistres sans réception de justificatifs.

1.2 Feature Engineering

1.2.1 Définition

Le Feature Engineering correspond à la création de nouvelles variables à partir de variables existantes. L'objectif est d'extraire de nouvelles caractéristiques en se basant sur celles présentes dans notre base de données. Les différents retraitements effectués sont basés sur des intuitions et des connaissances métiers afin de répondre aux mieux à la problématique.

1.2.2 Nouvelles variables

Cette partie présentera une liste non exhaustive des différentes variables créées à partir de celles existantes.

Retraitement de la variable date de naissance

La variable d'intérêt étudiée lors de ce mémoire est l'âge à la survenance du sinistre.

Retraitement de l'adresse pour obtenir le département

Cette étape a eu un rôle clef dans la modélisation pour établir une cartographie de la sinistralité. Elle a servi également lors de l'ajout de variables externes en tant que niveau de granularité choisi pour la jointure entre les variables externes et les variables internes.

Utilisation du numéro de sécurité sociale pour déterminer le sexe des individus

Chaque assuré possède un unique numéro de sécurité sociale qui lui est propre et qui sert à l'identification auprès de l'assurance maladie. L'information liée au sexe des assurés manquait dans les bases brutes extraites. Cette variable tient son importance dans le cadre de l'arrêt de travail, en effet on peut supposer que les hommes et les femmes n'ont pas la même sinistralité. Ne pas en tenir compte induirait une perte d'informations. En tenant compte de la décomposition d'un numéro de sécurité sociale présenté dans le tableau 1.1 il est possible de déterminer le sexe d'un individu.

Position	Interprétation
1	1 : Homme 2 : Femme
2-3	Deux derniers chiffres de l'année de naissance
4-5	Mois de naissance
6-7	Département de naissance
8 à 10	Code officiel de la commune
11 à 13	Numéro d'ordre de la naissance dans le mois et la commune
14-15	Clé de contrôle

TABLE 1.1 – Décompostion d'un numéro de sécurité sociale

La variable Sexe a ainsi pu être créée avec les modalités :
1 : chiffre correspondant aux hommes
2 : chiffre correspondant aux femmes

Nombre d'arrêt de travail par société

Le but de cette variable est d'arriver à déterminer un profil d'entreprise (chef d'entreprise) susceptible de participer à la fraude.

Création d'une variable liée à l'ancienneté dans le portefeuille au moment de l'arrêt

Elle est obtenue en faisant la différence entre la date de l'arrêt et la date d'effet du contrat.

Calcul de la franchise

La formule de la franchise choisie correspond à la différence entre le début de l'indemnisation et la date de déclaration de l'arrêt.

Censure et troncature

Il est important de définir ces notions afin de mieux appréhender les choix réalisés. La troncature à gauche est appliquée lorsque l'évènement à observer se produit avant la période d'étude, c'est-à-dire les arrêts ayant débuté avant la période d'observation. La base de données a été construite en sélectionnant les arrêts en fonction de leur date de survenance. Les arrêts composant notre échantillon concernent les survenances à partir du 1er janvier 2012, il n'y a donc pas de troncature à gauche.
La censure à droite consiste en une information partielle sur la durée de l'arrêt : certains arrêts ne sont pas observés jusqu'au bout, c'est-à-dire la sortie de l'état d'incapacité sera observée après la période d'observation. La figure 1.1 ci-dessous schématise les définitions faites précédemment

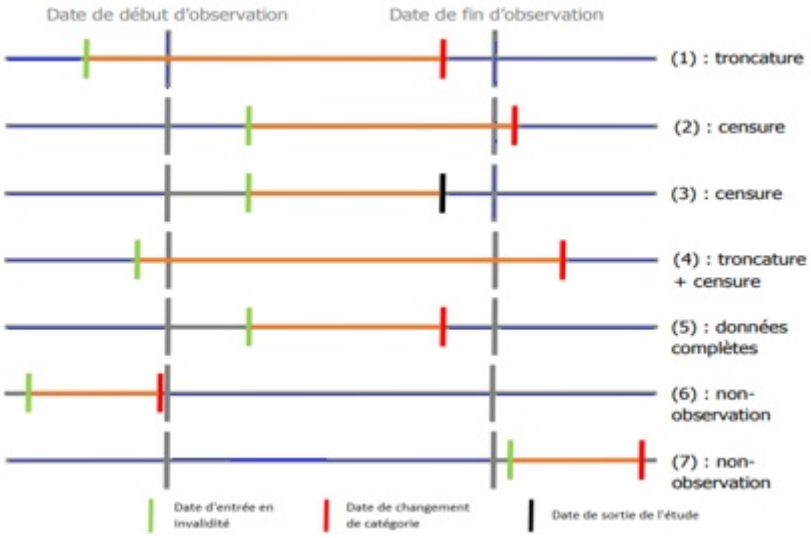


FIGURE 1.1 – Illustration de la troncature et de la censure

Dans le cadre de l'analyse effectuée dans ce mémoire le choix a été d'identifier comme censure les arrêts de travail sans motif de clôture et indemnisés durant les 3 derniers mois de la période d'observation (sinistres toujours en cours).

1.3 Rajout de variables externes

Pour enrichir la base de données, des données externes ont été récupérées. Elles proviennent principalement des données statistiques de l'INSEE.

1.3.1 Constat

Lorsque l'état de santé du salarié ne lui permet pas de travailler, il adresse l'arrêt de travail délivré par son médecin à son employeur et à la sécurité sociale dans les 2 jours. Les facteurs expliquant l'absentéisme peuvent donc être classés selon trois catégories :

— les caractéristiques associées à la santé et au bien-être,
— les attitudes et comportements associés à la motivation à être présent,
— les facteurs hors travail faisant obstacles à la présence.

1.3.2 Variables externes

Comme évoqué précédemment le risque arrêt de travail est marqué par une interdépendance entre la santé, le social et l'économique. Les variables externes choisies rajouteront un gain d'information dans ce sens. La majeure partie des données utilisées dans cette section sont les résultats d'études statistiques disponible sur le site de l'INSEE [1] et de la DREES [2]. L'INSEE préconise de manier avec précaution les effectifs inférieurs à 200 ou des zones de moins de 2 000 habitants la maille d'agrégation choisie est donc la maille départementale.

1.3.2.1 Variables liées au contexte économique

L'hypothèse faite est qu'il existe un lien entre l'arrêt de travail et la conjecture économique. En effet, on peut supposer que dans un contexte économique difficile la crainte de ne pas retrouver un emploi tend à freiner les salariés qui souhaitent quitter un emploi insatisfaisant, ce qui sur la durée augmente les arrêts qui sont utilisés comme facteur de l'absentéisme [3].

1. Institut national de la statistique et des études économiques
2. Direction de la Recherche, des Études, de l'Évaluation et des Statistiques
3. Le réseau Anact-Aract propose la définition suivante : « l'absentéisme caractérise toute absence qui aurait pu être évitée par une prévention suffisamment précoce des facteurs de dégradations des conditions de travail entendus au sens large : les ambiances physiques mais aussi l'organisation du travail, la qualité de la relation d'emploi, la conciliation des temps professionnel et privé, etc.

CSP

La catégorie socioprofessionnelle peut influencer la sinistralité également, en effet on peut supposer qu'un cadre aura moins tendance qu'un ouvrier à se mettre en arrêt ce dernier exerçant des tâches plus « pénibles ». Le jeu de données ne contient pas les informations relatives à la CSP des individus, pour contourner cela il a donc été rajouté 5 variables à savoir :

— la part des cadres,
— la part des professions intermédiaires,
— la part des employés,
— la part des ouvriers,
— une dernière catégorie regroupant toutes les autres modalités.

Ces variables renseignent le pourcentage par département pour chaque modalité en 2017.

Indice de pauvreté

Le taux de pauvreté correspond à la proportion d'individus dont le niveau de vie est inférieur pour une année donnée à un seuil, dénommé seuil de pauvreté (exprimé en euros). L'année de référence est 2017.

Niveau de vie

Il s'agit du niveau de vie par département en 2017. Le niveau de vie est égal au revenu disponible du ménage divisé par le nombre d'unités de consommation (uc). Le niveau de vie est donc le même pour tous les individus d'un même ménage. Les unités de consommation sont généralement calculées selon l'échelle d'équivalence dite de l'OCDE modifiée qui attribue 1 uc au premier adulte du ménage, 0,5 uc aux autres personnes de 14 ans ou plus et 0,3 uc aux enfants de moins de 14 ans.

1.3.2.2 Variables liées à la santé

L'arrêt maladie peut être considéré comme le signe d'une fragilisation de l'état de santé d'une personne. Intuitivement, on peut supposer l'existence d'un lien étroit entre la santé et les déclarations d'arrêts.

Personnel médical

Deux variables ont été retenues, les densités pour 100 000 habitants de l'ensemble des médecins et des généralistes.

Espérance de vie

Il s'agit de l'espérance de vie en fonction du sexe en 2019 par département.

Handicap

L'aide aux personnes dépendantes motive également les absences lorsque les salariés n'ont pas d'autres moyens pour venir en support. Trois variables ont dont été rajoutées à notre base de données :

— la proportion pour 1000 de personnes souffrant d'un handicap parmi les élèves (élémentaire et maternelle) en milieu scolaire ordinaire,
— la proportion pour 1000 de personnes souffrant d'un handicap parmi les élèves (collèges/lycées) en milieu scolaire ordinaire,
— la part des allocataires AAH (Allocation aux Adultes Handicapés) dans la population des 20 à 64 ans.

1.3.2.3 Variables liées à la démographie

Le choix s'est porté ici sur deux variables pouvant être considérées comme des indicateurs de la dépendance.

Indice de vieillissement

Cet indicateur se définit par la formule :

$$I_v = 100 \times (\frac{X_{65}}{X_{20}})$$

avec,
X_{65} : nombre de personnes de plus de 65ans.
X_{20} : nombre de personnes de moins de 20ans.

Taux de natalité

Il s'agit du taux de natalité par département en 2019.

1.3.2.4 Variables liées à la structure familiale

L'idée ici est de considérer que la structure familiale peut influencer le choix d'un assuré dans le cadre de l'arrêt de travail. Par exemple un parent peut choisir de se mettre en arrêt dans le but de passer du temps avec ses enfants. Le choix s'est porté sur les variables suivantes :

— le taux de divorce,
— le taux de famille dite " traditionnelle ",
— le taux de famille monoparentale ,
— le taux de famille recomposées,
— le nombre moyen de personne par ménage.

1.3.2.5 Bien-être et transport

La sensation de bien-être au travail peut être un frein à l'absentéisme. De même la distance entre le domicile et le travail peut avoir un impact assez fort sur l'absentéisme.

Ensoleillement

Il s'agit de la mesure du nombre d'heures d'ensoleillement en 2018 par département.

Moyen de transport sur la distance domicile travail

Quatre variables ont été utilisées :

- taux de personne se déplaçant à la marche,
- taux de personne utilisant les transports en commun,
- taux de personne utilisant un moyen de transport personnel (deux roues, véhicules...),
- taux de personne travaillant à leur domicile.

Au total notre base a été enrichie de 25 variables. La base finale étant à disposition nous pouvons procéder à l'établissement des statistiques descriptives.

Chapitre 2

Analyse exploratoire de la base

Le but de cette section est de mieux connaître la base à disposition. Cette étape est essentielle dans le processus de modélisation afin d'appréhender au mieux la structure des données.

2.1 Analyse de la forme

Après les différents retraitements mentionnés plus haut notre base final contient 41910 observations et 48 variables (voir annexe 4 pour la liste complète des variables, leurs typologies et descriptions associées)

2.1.1 Typologie des variables

La typologie des variables est la suivante :

- 38 variables numériques,
- 9 variables qualitatives,
- 1 variable au format Date.

2.1.2 Valeurs manquantes

La figure 2.1 ci-dessous est une image de la base à disposition.

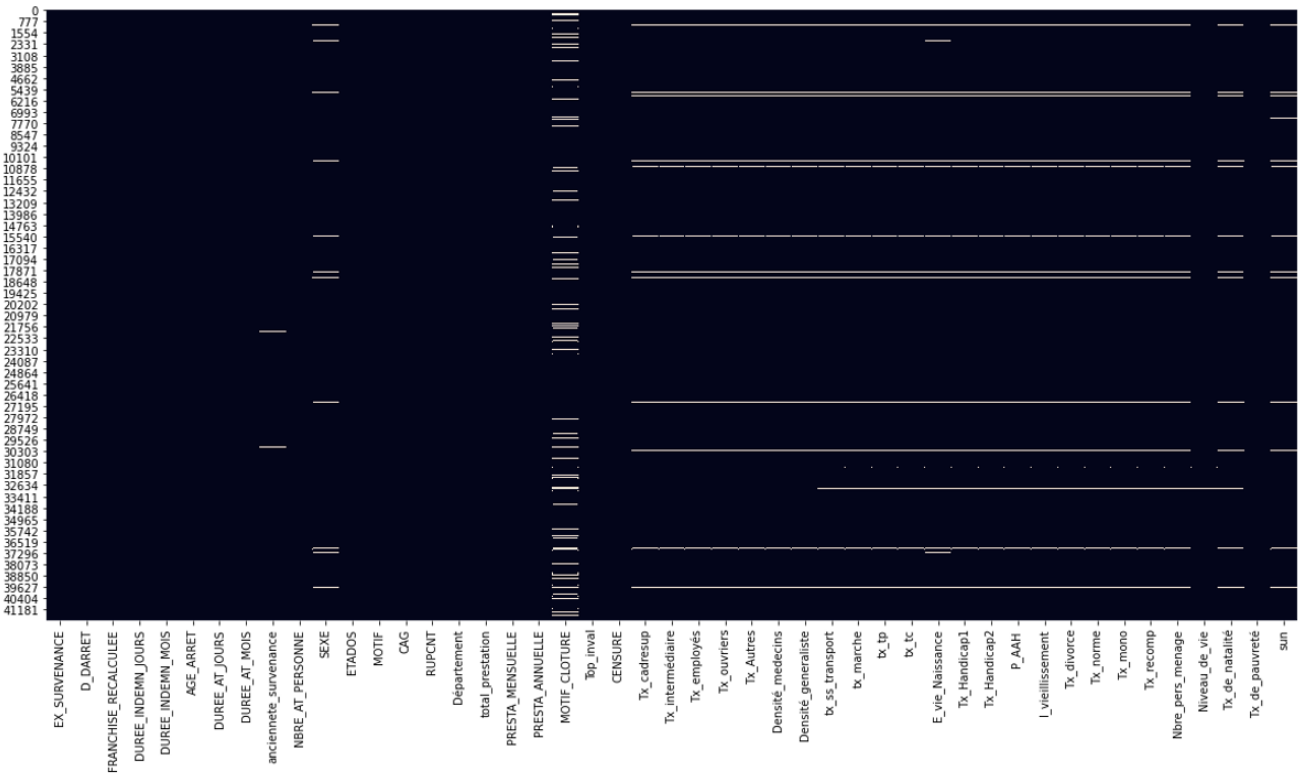


FIGURE 2.1 – Représentation de l'architecture de la base en fonction des valeurs manquantes

Les valeurs en blanc correspondent aux valeurs manquantes dans le jeu de données. La variable liée au motif de clôture semble être celle qui présente le plus de valeurs manquantes.

Ce graphique peut être subdivisé en deux :

— une partie gauche (variables internes) présentant très peu de variables manquantes,
— une partie droite (variables externes) présentant une série de valeur manquantes sur des observations en particulier.

Cela s'explique par la fait que plusieurs statistiques correspondant à certains départements (la corse et l'outre-mer notamment) étaient manquantes dans les données récupérées sur le site de l'INSEE.

2.2 Analyse du fond

2.2.1 Focus sur l'âge et le sexe

Dans la base de données mise à disposition, le pourcentage des femmes est de 55% tandis que celui des hommes est de 45%.

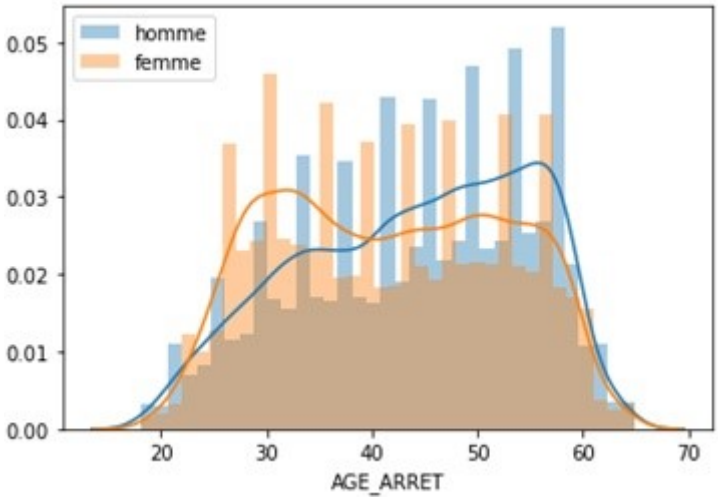


FIGURE 2.2 – Histogramme de la variable âge en fonction du sexe

Peu d'arrêts sont observés aux extrémités, le départ à la retraite où le fait que ces salariés soient encore étudiants peuvent être les raisons de ces observations.
La figure montre une différence significative entre les sexes. La distribution des femmes est stable entre 40 et 60 ans tandis que celle des hommes est en constante évolution jusqu'à atteindre son pic autour des 60ans. Chez les femmes le pic est atteint autour de 30 ans cela peut traduire des arrêts liés à la maternité.

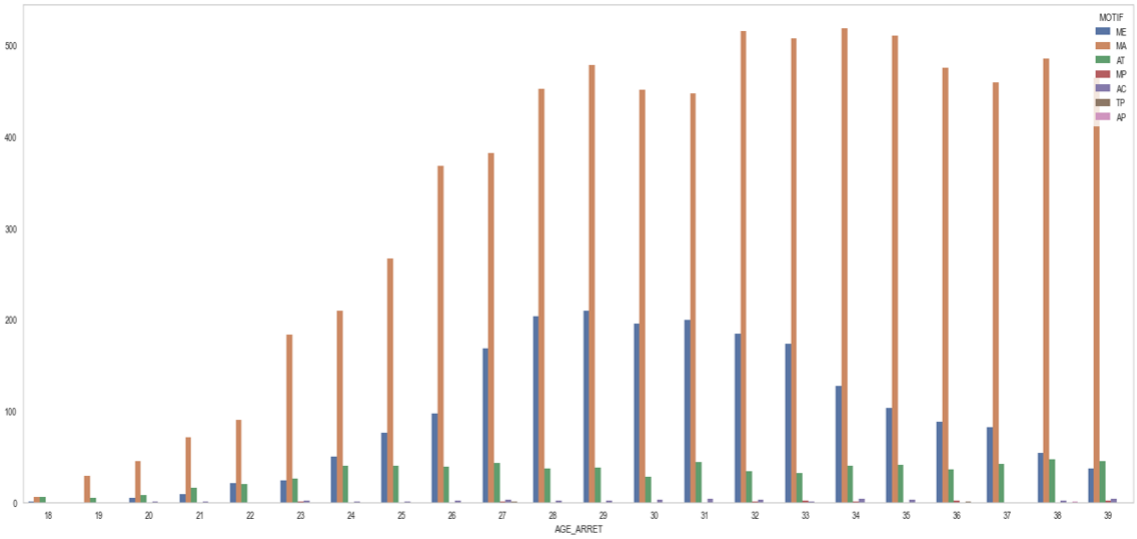


FIGURE 2.3 – Distribution des arrêts de travail de sexe Féminin entre 18 et 39ans

La figure 2.3 confirme les hypothèses sur les raisons des arrêts dans la population de sexe féminin. Bien que la majeure partie des arrêts soit déclarée pour raison de maladie (en orange) les motifs de maternité (en bleu) sont concentrés autour des 30 ans.

Au vu de ces analyses la variable âge à l'arrêt sera discrétisée de la fonction suivante :

Groupe	Tranche d'âge
0	<30ans
1	[30,40[
2	[40,60]
3	>60

TABLE 2.1 – Discrétisation de la variable âge

2.2.2 Lien entre durée en arrêt, le motif et le sexe

Le graphique ci-dessous présente la durée en arrêt (en mois) par sexe.

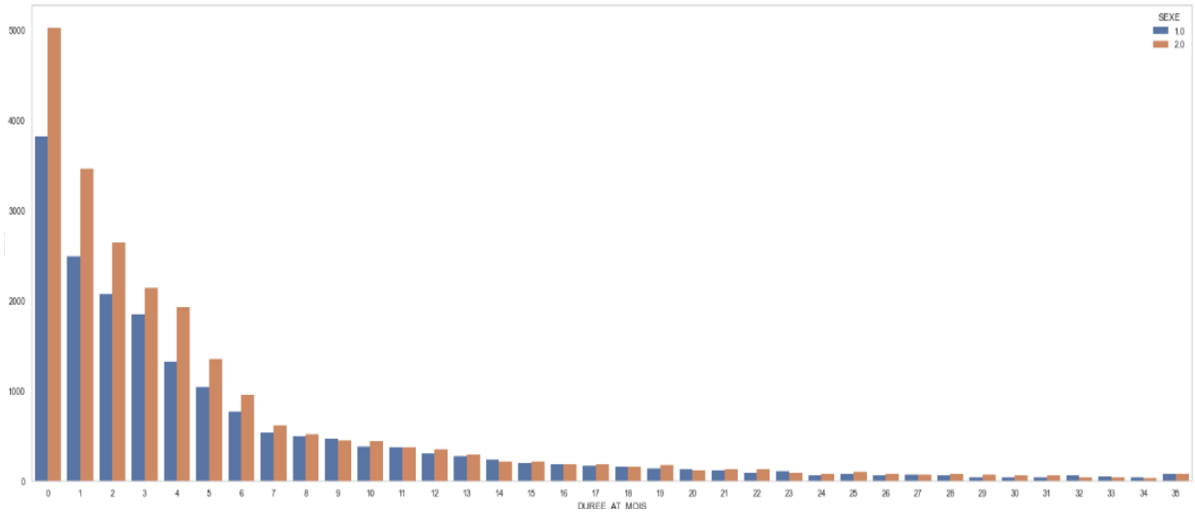


FIGURE 2.4 – Durée de l'arrêt en fonction du sexe

La première déduction faite de ce graphique est que la majeure partie des arrêts ont une durée inférieure à 6 mois.

Pour ces arrêts une différence semble se dégager entre les hommes (en bleu) et les femmes (en orange).

Existe-il une différence de comportement entre les hommes et les femmes en fonction du motif de l'arrêt ? Le Tableau ci-dessous apportent certains éléments de réponse à cette question.

Motif	Sexe	Durée moyenne de l'arrêt (en jours)
Accident	Homme	147
	Femme	147
Accident causé par un tiers	Homme	215
	Femme	327
Accident du travail	Homme	140
	Femme	136
Maladie	Homme	191
	Femme	177
Maternité	Homme	12
	Femme	127
Maladie Professionnelle	Homme	304
	Femme	300

TABLE 2.2 – Nombre de jours moyen en arrêt de travail par sexe

Il n'y a aucune différence significative entre les sexes en observant les motifs " Maladie Professionnelle " et "Accident". Cependant sur le motif "Accident causé par un tiers" une réelle différence existe. Il en est de même pour la maternité la durée moyenne en jour est significativement différente entre les hommes et les femmes.

2.2.3 Focus sur la zone

La zone d'habitation peut avoir un rôle clef dans la modélisation. La figure ci-dessous nous présente le nombre d'arrêts déclarés par département.

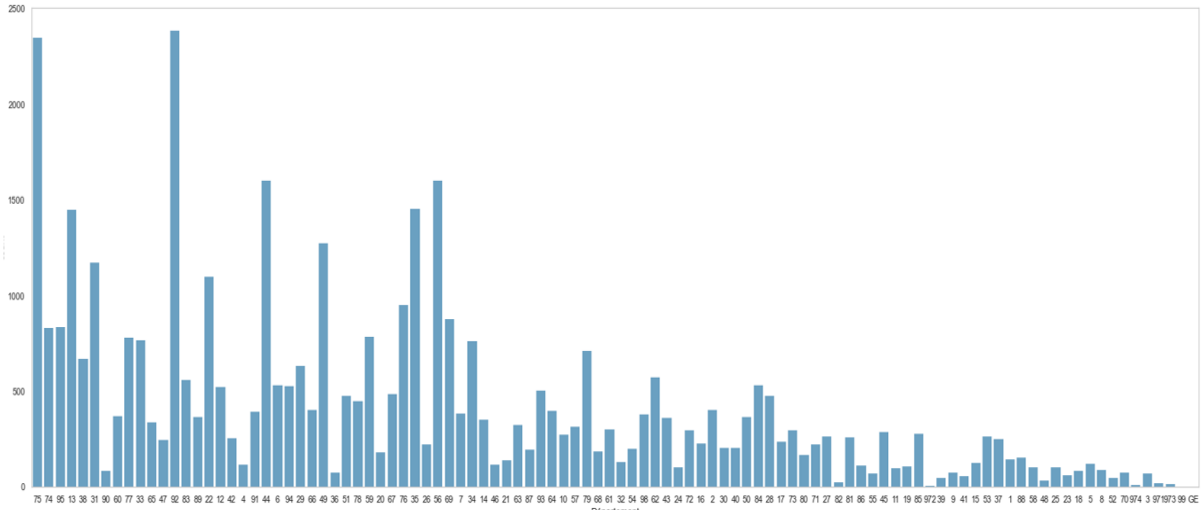


FIGURE 2.5 – Nombre d'arrêts de travail déclarés en fonction du département

Deux départements d'Ile-De-France semblent se distinguer des autres, il s'agit des hauts de seine (92) et de Paris (75). A ce stade, il est difficile de faire une quelconque hypothèse sur le lien entre l'arrêt de travail et la zone d'habitation en effet l'information concernant le nombre d'assurés par département est manquante.

2.2.4 Lien entre invalidité et l'âge

La figure 2.6 montre le croisement entre l'âge et l'entrée en invalidité.

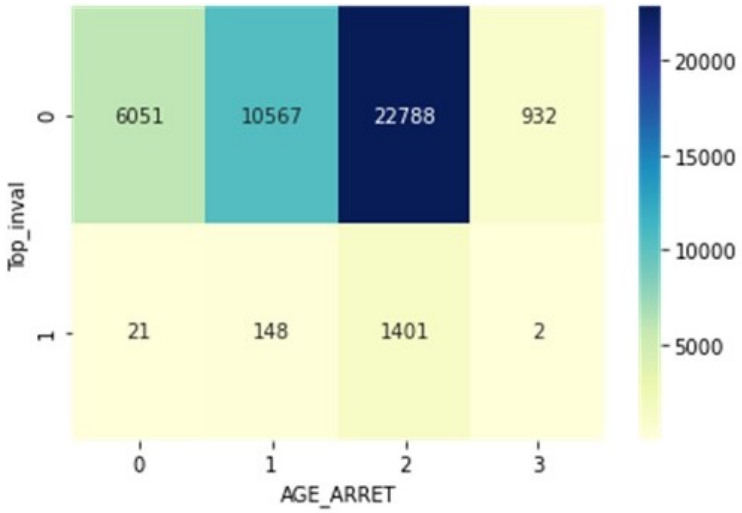


FIGURE 2.6 – Tableau croisé entre la variable Top Invalidité et la variable âge

L'analyse révèle que les individus dans la classe 2 à savoir les [40,60[sont ceux dont le taux de passage en invalidité est le plus élevé.

2.2.5 Saisonnalité

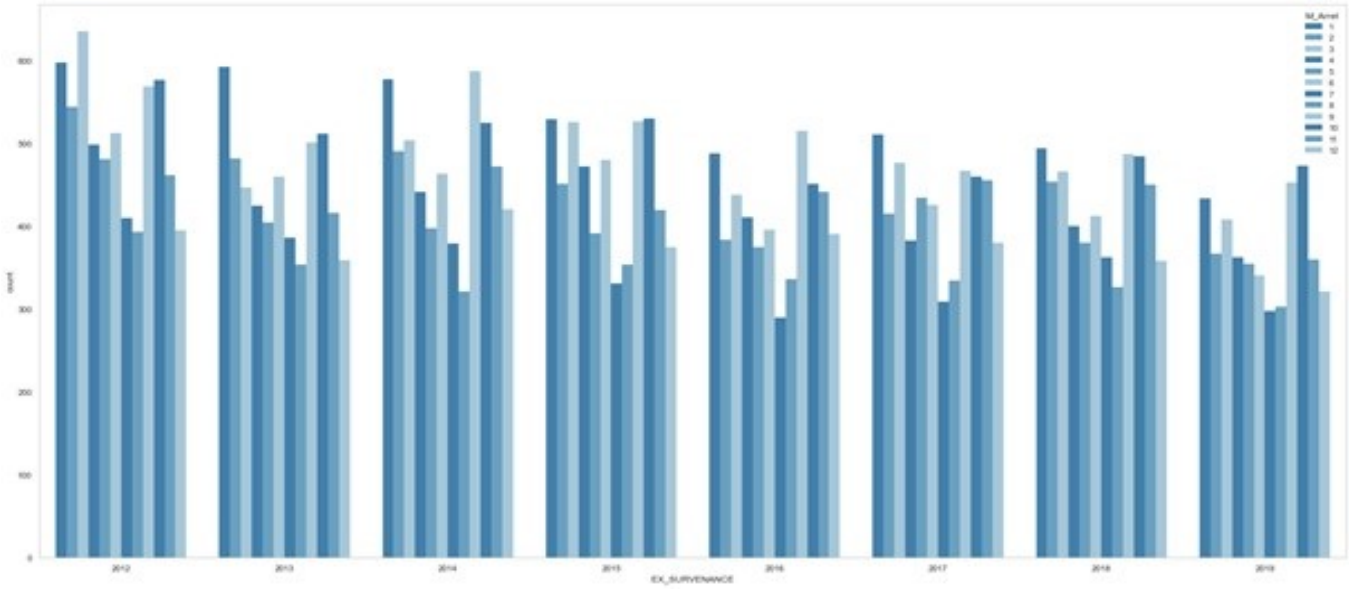


FIGURE 2.7 – Sinistralité par an en fonction des 12 mois de l'année

Une saisonnalité semble se dégager en observant le graphique. En effet pour les 8 années observées le pic d'observation se situe sur les 3 premiers mois de l'année (Janvier-Février-Mars) ou sur les 3 mois (Septembre-Octobre et Novembre). De plus les mois de Juillet et Août (été) semblent être les mois les moins touchés par les déclarations, marquant ainsi la rupture avec le pic observé en Septembre.

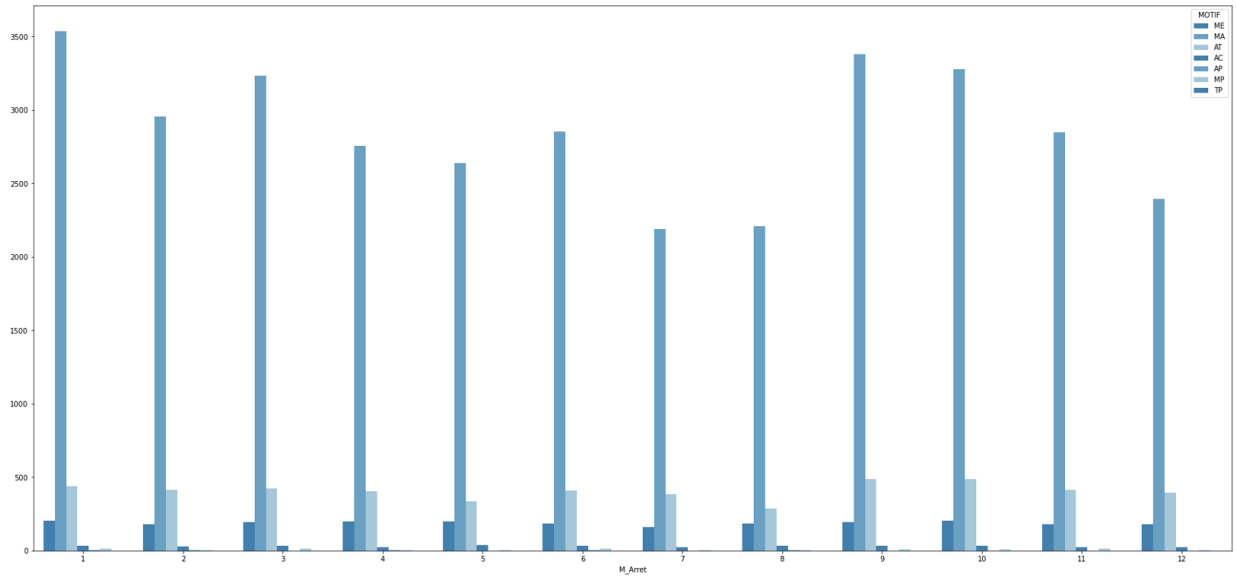


FIGURE 2.8 – Les motifs d'arrêts par mois

La figure 2.8 semble apporter un élément de réponse sur la saisonnalité observée. Les motifs de déclarations d'arrêt de travail semble tous stables en fonction des mois, à l'exception des arrêts pour cause de maladies. Les hausses observées sur ces valeurs correspondent à celles des mois ayant un pic de déclarations. En tenant compte cette analyse, la variable mois a été catégorisée en :

— 0 pour les mois ayant une sinistralité faible (en général température moyenne plus élevée que les autres)
— 1 pour les autres

Tout au long de ce chapitre nous avons pris connaissance de notre base au travers de statistiques et d'analyses qui ont permis de réaliser certains retraitements légers pour ne pas tomber dans les pratiques liées à l'apprentissage supervisée. Les principaux enseignements tirés de ses statistiques sont :

— la sinistralité évolue différemment lorsqu'on croise les variables sexe et âge,
— une saisonnalité est observée causée principalement par les arrêts pour cause maladie.

Les données étant au format propice, le chapitre suivant traitera de la modélisation.

Chapitre 3

Modélisation

L'objet de ce chapitre est la classification des données en 2 classes (observations normales et anormales) via une approche non-supervisée. Cette approche consiste à développer un modèle d'apprentissage statistique permettant de regrouper les données en différentes classes non connues à priori. L'apprentissage non supervisé ne présuppose aucun étiquetage préalable des données.
L'objectif est que le modèle analyse la structure des données et classe les observations selon leur degré de similitude. L'idée n'est pas de déterminer des règles permettant de prédire l'appartenance à une classe cible, mais à identifier des règles de regroupement au sein de différentes classes définies par le modèle lui-même.
Ceci est fait en optimisant les critères de regroupement des individus dans des classes, chacune la plus homogène possible et, entre elles, les plus distinctes possible. D'un point de vue schématique, l'approche non supervisée revient à laisser le modèle analyser les données sans lui préciser ce qu'il doit trouver.

L'application de l'apprentissage non-supervisé à la détection d'anomalies revient à identifier dans la base de données les échantillons dont les caractéristiques sont assez différentes des autres.
L'algorithme testé dans le cadre de ce mémoire est l'Isolation Forest, un algorithme basé sur les arbres de décision.

3.1 Rappels théoriques du modèle utilisé

Cette section a pour but de présenter les notions théoriques de l'algorithme d'Isolation Forest afin de mieux appréhender les analyses qui seront réalisées une fois ce modèle implémenté.

Les notions détaillées proviennent de l'article publié par Fei Tony Liu & Kai Ming Ting [12] .
C'est une technique innovante dans la détection d'anomalies en ce sens où la plupart des modèles existant établissent d'abord une typologie des observations dite « normale » et par la suite désigne comme « anomalies » les profils qui ne sont pas conformes à la typologie faite en premier lieu.

Les 2 inconvénients majeurs de ces anciennes méthodes sont :

— le détecteur est optimisé pour établir un profil des observations dite «normale » ce qui peut avoir pour conséquence un très faible nombre d'anomalies détectés ou des faux positifs (instances normales catégorisées comme anormales),

— ces méthodes ne sont pas applicables sur des grandes bases de données en raison d'une grande complexité de calcul.

L'algorithme d'isolation Forest permet de remédier à ces inconvénients.
L'idée de cet algorithme est de partitionner le jeu de données en choisissant au hasard un paramètre (variable) puis de réaliser des découpages par le choix arbitraire d'une valeur entre les valeurs maximale et minimale.
L'hypothèse faite est de considérer que plus le nombre de découpages avant d'isoler un point est faible, plus il est fort probable que ce point soit une anomalie comme nous pouvons le voir sur la figure ci-dessous en deux dimensions (2 variables).

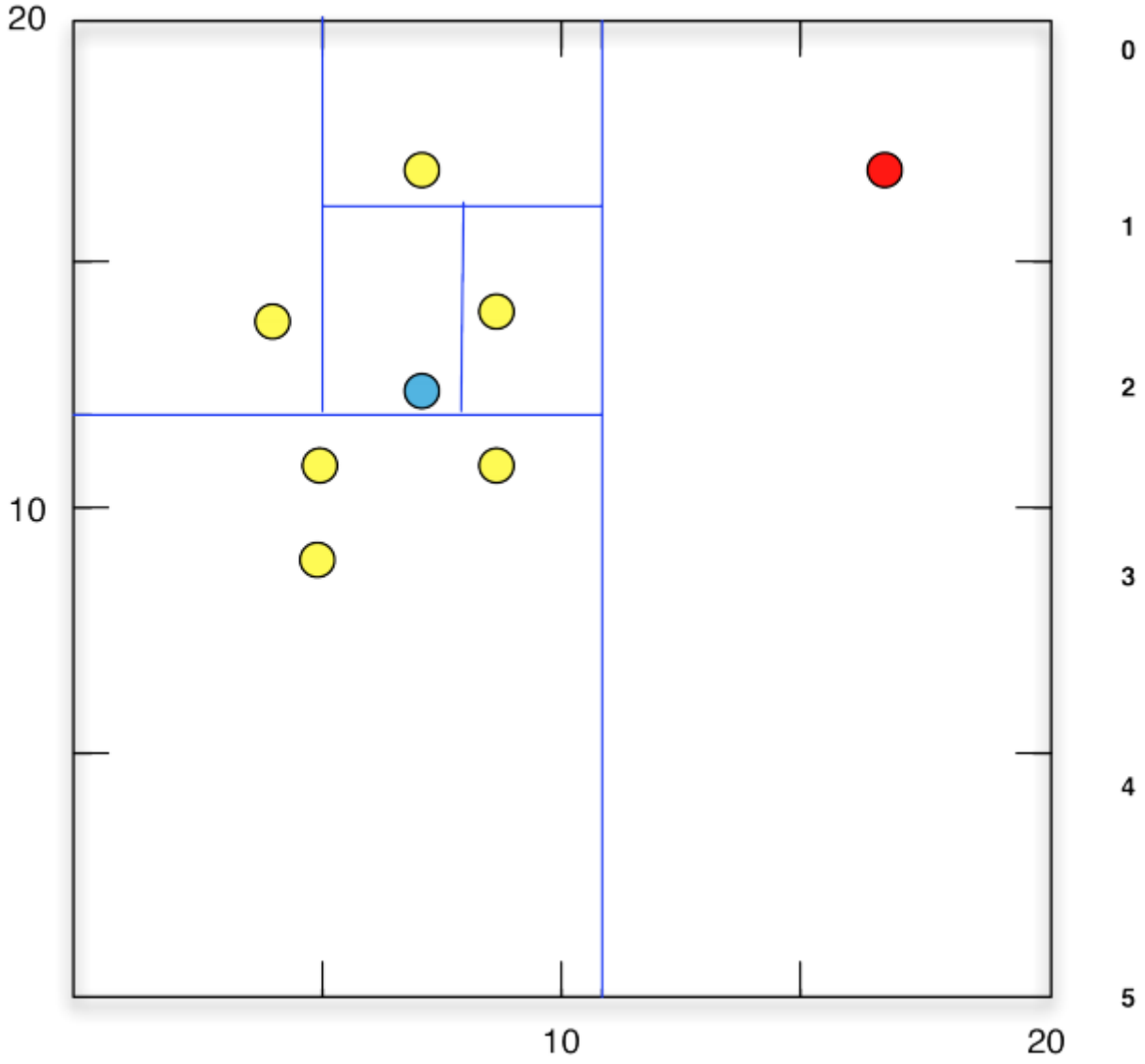


FIGURE 3.1 – Segmentation en deux dimensions

En effet le point rouge (point atypique) semble isolable plus facilement que les autres points dans la masse comme le point bleu. Ces découpages s'inspirent des forêts aléatoires, les partitions itératives permettent de construire un arbre binaire où chaque nœud possède exactement zéro ou deux fils comme illustré sur la figure ci-dessous.

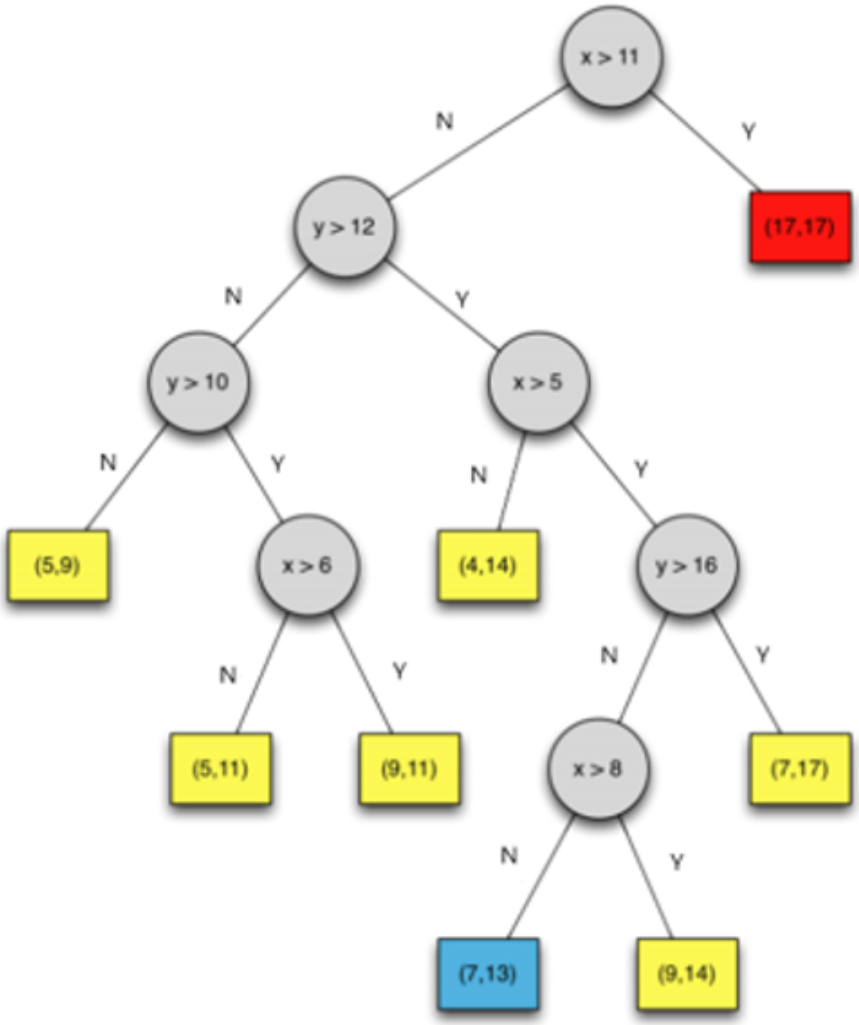


FIGURE 3.2 – Exemple d'une isolation tree, cas bidimensionnel

La probabilité d'isoler un échantillon dit « anormal » via des découpages est bien plus grande que celle d'un point se trouvant dans la masse (observation normale). Néanmoins, bien que la probabilité soit faible, il est possible d'isoler un échantillon normal (dans la masse de données) avant d'isoler une anomalie.

Pour pallier à cet inconvénient plusieurs séquences d'isolation sont réalisées et le résultat final est égale à la moyenne de toutes les estimations, l'Isolation Forest fait donc partie des techniques d'ensemble de générateur de type « arbre ».

Implémentation de l'algorithme

La mise en place de l'algorithme se fait en deux étapes :

- une étape d'apprentissage,
- une étape de prédiction.

Etape d'apprentissage

Etape 1 : Sélection d'un échantillon et construction d'un algorithme d'isolation sur ce dernier.
Etape 2 : Répétition de l'étape 1 pour construire un ensemble d'algorithme d'isolation.
Rappel : la construction de l'arbre consiste en la sélection d'une variable au hasard et par le choix d'une valeur de partition ensuite.

Etape de prédiction

Etape 1 : Calcul d'un score d'anomalies

$$S(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

avec,
x : une observation de notre données (une ligne).
$h(x)$: la longueur du chemin de l'observation x c'est-à-dire le nombre total de séparations nécessaires pour isoler l'observation x.
$E(h(x))$: correspond à la moyenne sur tous les abres de h(x).
n : correspond à la taille des échantillons dans les différents arbres construits.
$c(n)$: la valeur moyenne des longueurs de chemin sur les n différents arbres.
Cette valeur peut être estimée par :

$$c(n) = 2H(n-1) - (2\frac{(n-1)}{n})$$

où le terme H(n-1) peut être estimé par la constante d'Euler ln(n-1)+ 0.5772156649.

On en déduit de l'équation $S(x,n)$ les conclusions suivantes :

- $E(h(x)) << c(n)$ alors $S(x,n) -> 1$ l'observation x est probablement une anomalie,
- $E(h(x)) -> c(n)$ correspond aux échantilons ne présentant pas d'anomalies distinctes,
- les observations présentant un score inférieur à 0,5 sont celles se trouvant à l'intérieur de la masse de données.

En considérant $s \in]0\ ,1]$ et $h(x) \in]0\ ,\ \text{n-1}]$, la figure ci-dessous illustre la relation entre $h(x)$,$c(n)$ et s.

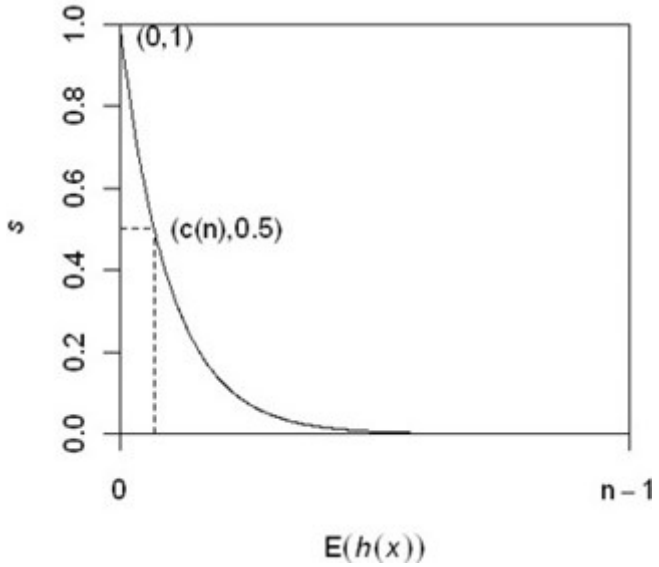


FIGURE 3.3 – Evolution du score d'anomalie en fonction de E(h(x)), source[12]

3.2 Application de l'algorithme

3.2.1 Preprocessing

L'objectif de cette étape est la mise en forme de nos données dans le format propice pour les intégrer aux algorithmes. l'Isolation Forest n'utilisant pas la notion de distance aucune standardisation n'est nécessaire. Cependant il est nécessaire de faire de l'encodage afin de transformer les variables qualitatives en variables numériques.

3.2.2 Modélisation

L'algorithme d'Isolation Forest a été implémenté avec la commande IsolationForest du module sklearn.ensemble.
Plusieurs paramètres sont nécessaires pour l'implémentation dans Python dont le taux de contamination. Il s'agit du taux d'anomalies à détecter donc d'observations anormales. Etant dans un cadre non-supervisé, cette valeur est soumise à l'avis d'expert. Pour contourner ce problème la procédure suivie est la suivante :

Etapes de la mise en place de l'algorithme

1. **Construction d'un score d'anomalie pour chaque observation.**
2. **Tracé de l'histogramme des scores d'anomalie**
3. **Par l'analyse graphique de l'histogramme des anomalies, choix d'une règle de décision**
   a) Choix d'un seuil de décision.
   b) Si le score d'une observation est supérieur à ce seuil il s'agit probablement d'une anomalie.

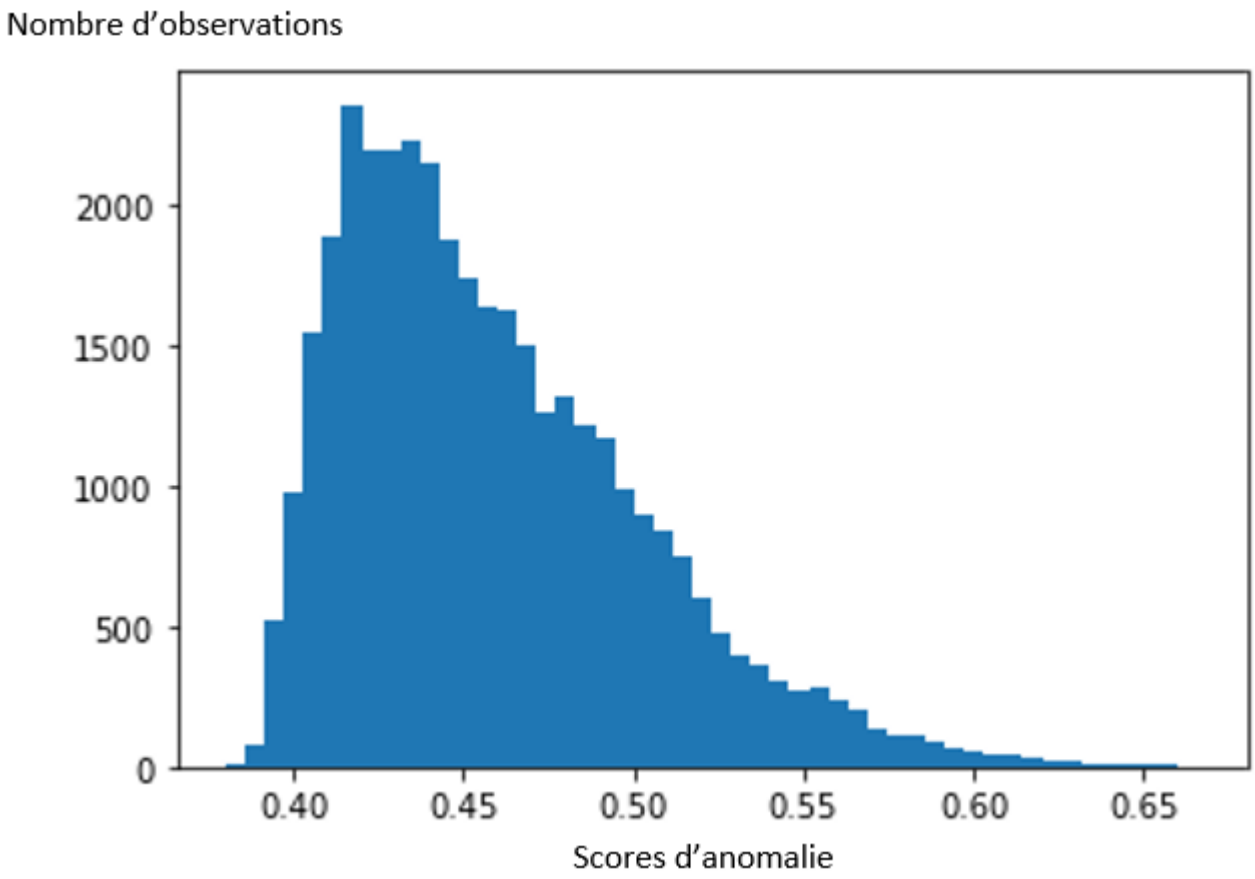


FIGURE 3.4 – Histogramme des scores d'anomalie

La figure ci-dessus représente l'histogramme des scores d'anomalie enregistrés pour chaque observation de la base de données. Pour rappel, les observations « atypiques » ont 2 caractéristiques à savoir :

— elles représentent la classe minoritaire,

— leurs scores d'anomalie sont élevés par rapport à la moyenne.

Compte tenu de ces observations et de l'histogramme le seuil choisi est 60%. Les observations anormales sur ce critère représentent en proportion 0,5% de notre base de données.

3.3 Présentation des résultats

Dans le tableau suivant sont consignés les résultats des 3 modalités majoritaires de différentes variables en fonction de la base observée.

	Données Normales	Données atypiques
Durée AT en mois	0 : 23%	35 : 15%
	1 : 15%	33 : 8 %
	2 : 12%	25 : 5%
Age à l'arrêt	[40,60] : 57%	[40,60] : 77%
	[30,40[: 26%	[30,40[: 15%
	>60 : 15%	<30 : 4%
Motif	Maladie : 82%	Maladie : 90%
	Accident du travail : 12%	Accident du travail : 2%
	Maternité : 5%	Accident causé par un tiers : 1%
Mois Arrêt	Janvier : 11%	Janvier : 15%
	Mars : 10%	Septembre : 10%
	Septembre : 9%	Décembre : 9%

FIGURE 3.5 – Présentation des résultats des clusters pour différentes variables

L'analyse variable par variable fait ressortir deux enseignements :

— les données normales sont assez proches de la base de départ ce qui s'explique par le fait qu'elles sont majoritaires donc assez proche de la base de données intiale,

— pour la variable "Durée AT en mois" l'algorithme semble avoir ciblé les arrêts minoritaires (de longue durée) dans la base.

L'analyse variable par variable ne nous permet pas de comprendre les différences entre les observations topés comme « anomalies » et celles considérées comme « normales ». Une analyse plus globale permet de ressortir certaines tendances :

— les arrêts de longue durée pour motif de maternité déclarés par les hommes sont observés uniquement dans la base des observations atypiques,

— majoritairement (dans la base des observations normales) les arrêts observés dans les mois hivernaux ne concernent pas les personnes de moins de 30 ans. Cependant dans la

base des observations atypiques sur la même période des arrêts longue durée sont observés majoritairement par des personnes de sexe féminin hors motif de maternité.

— Les arrêts observés sur les mois estivaux sont globalement de courte durée en comparaison des autres sur la base des observations normales. Néanmoins, la base des anomalies met en évidence des arrêts plutôt de longue durée et dont les motifs sont souvent liés aux accidents (Accident de travail, Accident de travail causé par tiers personne).

L'analyse croisée des variables nous a permis d'établir un portrait-robot des observations considérées comme atypiques, la section suivante va traiter de l'évaluation de ces clusters.

3.4 Procédure d'évaluation

Dans un cadre non-supervisé, le back-testing représente le challenge majeur. Pour rappel l'objectif de l'apprentissage non-supervisé est d'analyser la structure des données sans étiquettes afin de réaliser certaines tâches telle que la classification. L'absence d'étiquettes ne permet donc pas de juger de la pertinence des différents clusters réalisés.
Le but de cette section est de challenger le modèle précédemment utilisé par des moyens détournés.
L'approche considérée ici est une approche visuelle.
La base contient 47 variables utilisés pour entraîner l'Isolation Forest. Dans cette section, sera présenté la visualisation des résultats de la classification en 2 dimensions.
Pour atteindre cet objectif, les données seront dans un premier temps standardisées, puis représentées en 2 dimensions via la technique TSNE .

TSNE

D'un point de vue mathématique l'Analyse en Composantes Principales peut être considérée comme une méthode de projection qui permet de projeter les observations depuis un espace à p dimensions (p variables) vers un espace à k dimensions ($k < p$) avec pour but qu'un maximum d'information soit conservée (l'information est ici mesurée au travers de la variance totale du nuage de points) sur les premières dimensions. L'objectif de cet algorithme est donc de simplifier la structure des données tout en conservant les principales informations. Cette technique permet d'accélérer l'apprentissage de certains algorithmes et de lutter contre le fléau de la dimension.
La TNSE l'abréviation de " t-Distributed Stochastic Neighbor Embedding" est une technique présentée par Laurens van der Maaten et Geoffrey Hinton [13] . Contrairement à l'analyse en composantes principales elle permet la réduction de dimension dans le cas où il n'existe pas de relation linéaire entre les variables.

Considérons x_i et x_j deux points, on construit une gaussienne autour de x_i(centrée en x_i et de variance σ_i). La probabilité conditionnelle pour x_i d'être choisit comme voisin de x_j.

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)}$$

On en déduit la mesure de similarité entre les deux points comme la probabilité jointe p_{ij} définit par :

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

Toutes les dimensions seront projetées sur un espace de dimension inférieur, et les différentes probabilités seront calculées à nouveau suivant le même principe.
La probabilité jointe dans cet espace réduit q_{ij} entre y_i et y_j est donnée par :

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k\neq i}(1 + ||y_i - y_k||^2)^{-1}}$$

La différence de probabilité ici est que la distribution utilisée en dimension réduite est celle de student à 1 degré de liberté.

Pourquoi ce changement de loi ?

Pour comprendre aisément imaginons 3 points dans un espace à 2 dimensions.

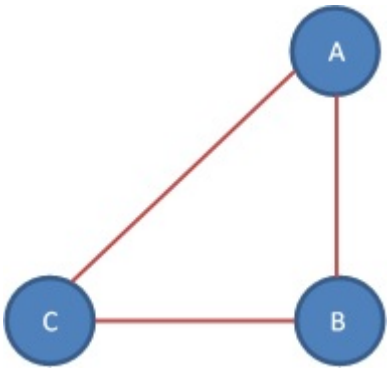


FIGURE 3.6 – Projection de 3 points dans un espace 2d

En essayant de réduire la dimension l'espace à une dimension tout en conservant la logique des rapports de distance($d(a,b) < d(a,c)$) on aurait le schéma suivant :

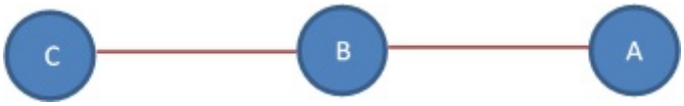


FIGURE 3.7 – Réduction de l'espace en une dimension

La distance entre les points C et B est conservée mais celle entre C et A a fortement augmenté. Ce phénomène arrive très souvent dans la réduction de dimension d'où l'utilisation d'une loi

ayant une queue plus lourde que la gaussienne afin d'atténuer ce phénomène.
La construction de l'espace réduit est réalisée de manière à ce que la similarité entre les points x_i et x_j soit correctement reproduite dans l'espace de plus faible dimension par les points y_i et y_j.
Cela revient donc à réduire les dissimilarités entre les distributions P et Q.
La mesure de divergence choisie est celle de Kullback–Leibler (KL) présentée en 5.3.
La fonction de coût qui permet d'évaluer la projection dans un espace réduit est donc C :

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

La minimisation de cette fonction est réalisée par la descente de gradient , la forme de la dérivée est donnée par :

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Présentation du jeu de données en 2 dimensions

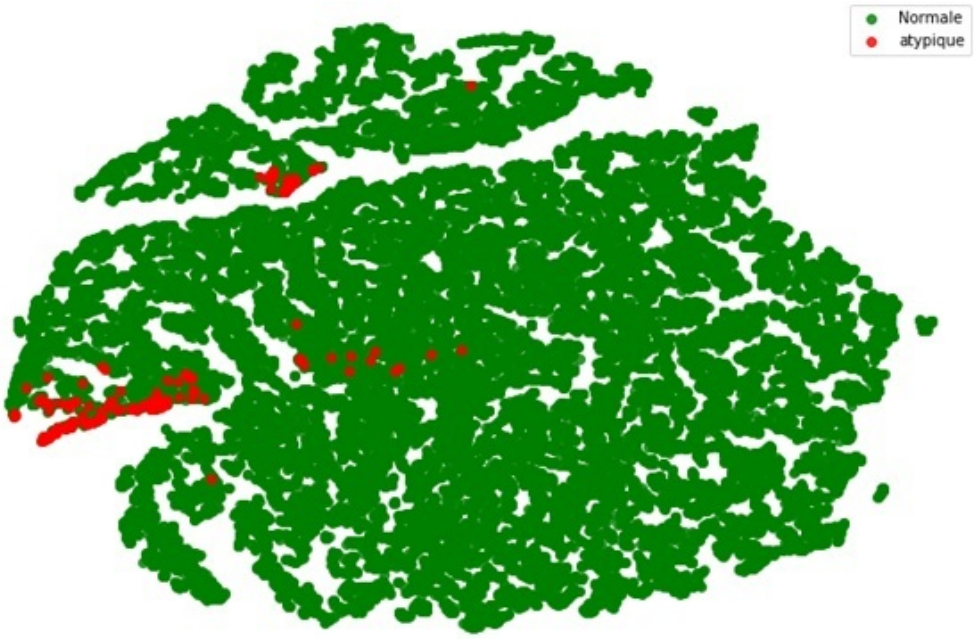


FIGURE 3.8 – Représentation en deux dimensions des clusters réalisés via l'algorithme d'Isolation Forest

Les clusters ne semblent pas correctement identifiés et délimités dans l'espace en 2 dimensions. Le résultat serait peut-être différent si la projection était réalisée dans un espace supérieur à 2. A ce stade de l'analyse il est difficile de se prononcer sur la performance de l'algorithme.

3.5 Conclusion

Le manque de données étiquetées a aiguillé l'étude dans un cadre non-supervisé. L'apprentissage non-supervisé revient à laisser le modèle analyser la structure des données sans lui préciser ce qu'il doit trouver pour ensuite réaliser certaines tâches telles que la classification dans notre cas. Cette approche a permis de mettre en évidence les observations présentant une structure atypique au sein de la base de données.
A la problématique de justifier de la pertinence des clusters réalisés elle peut se mesurer au travers du gain d'information que chacune des classes apporterait sur la base.
Prenons le cas où les observations atypiques s'avèreraient être des cas de fraudes cela justifierait de la pertinence de la segmentation effectuée.
Nous ne disposons pas de données sur la fraude en effet les assureurs sont peu enclins à partager ce genre de données.
Le choix a donc été de s'orienter vers une base en libre-service disponible sur Kaggle sur la problématique de détection de fraude à la carte bancaire.

Troisième partie

Application à la détection de fraudes

La deuxième partie de ce mémoire a été consacrée à la détection d'observations atypiques dans un cadre non-supervisé. L'avantage principale de cette approche est la facilité de sa mise en place, cependant l'inconvénient majeur reste la difficulté à tester la pertinence des modèles obtenus. L'objet de cette troisième et dernière partie est donc de tester la robustesse de l'algorithme d'Isolation Forest dans la détection d'anomalies.
Pour ce faire, nous utiliserons une base de données en libre accès contenant des données étiquetées comme frauduleuses ou non-frauduleuses.
En effet, il est naturel de considérer qu'une fraude révèle d'un caractère anormal. L'Isolation Forest sera implémenté puis testé d'une part afin de mesurer son pouvoir de prédiction et d'autre part afin de le comparer aux algorithmes d'apprentissage supervisé.

Un mot sur la fraude en Assurance

Il n'existe pas de définition légale de la fraude. Néanmoins, l'ALFA [1] définit la fraude comme : « ***Un acte intentionnel, réalisé par une personne morale ou physique, afin d'obtenir indûment un profit du contrat d'assurance*** ».
Cela implique donc une mauvaise foi de l'assuré. Il n'y a pas de fraude lorsqu'il s'agit d'une erreur ou lorsque l'assuré agit dans la croyance légitime de ne pas réaliser d'irrégularité. La fraude à l'assurance se définit donc comme un acte délibéré commis contre un assureur ou un intermédiaire d'assurance dans le but d'obtenir un avantage qu'il soit financier ou non. Il est à noter que près d'un demi-milliard d'euros de fraudes ont été identifiée en 2018 uniquement en IARD selon l'ALFA.
Selon un sondage mené par l'institut CSA, paru en Novembre 2019, plus d'un Français sur 10 admet avoir déjà effectué une fraude à l'assurance et près de 30% affirment qu'ils pourraient le faire à l'avenir s'ils en avaient l'occasion.
Ces fausses déclarations engendrent d'énormes pertes pour les compagnies d'assurance. Elles se présentent sous plusieurs aspects, pour en citer quelques-unes :

— fraude à la souscription : il peut s'agir de fausses déclarations,
— fraude lors de la déclaration des sinistres,
— incident provoqué intentionnellement.

Prenons le cas des fraudes en arrêt de travail. La fraude dans ce cas représente toute situation où l'assuré est en arrêt de travail alors que ce n'est pas justifié, c'est-à-dire qu'il a la capacité de travailler.
La fraude peut intervenir tout au long de la vie de l'arrêt. Cet arrêt de travail peut être frauduleux dès l'origine, si l'assuré n'a jamais été malade, mais il peut également le devenir dans le cas où la personne a été malade et ne l'est plus tout en déclarant être tout de même en arrêt.

1. l'Agence pour la Lutte contre la Fraude à l'assurance (ALFA)est une association loi de 1901, sans but lucratif, dont l'objet est de promouvoir la lutte contre la fraude dans le secteur de l'assurance

Chapitre 4

Etude de la base à disposition

4.1 Présentation de la base

Par soucis de confidentialité, les assureurs sont peu enclins à partager leurs données d'assurance sur la fraude même anonymisées.
Dans le cadre de ce mémoire, l'étude portera donc sur une problématique analogue : la détection de fraudes sur les cartes de crédit.

La base étudiée contient les transactions effectuées par des détenteurs de cartes de crédit Européenne en Septembre 2013 en Europe. Une fenêtre d'observation de 2 jours a été choisie pour l'observation de ces transactions.
Le jeu de données contient 284807 lignes représentant chacune les transactions, 30 colonnes qui sont toutes numériques et aucunes valeurs manquantes. Pour des raisons de confidentialité, le jeu de données originale a été transformé, la décomposition des variables est la suivante :

- 28 variables V1 à V28 qui sont les 28 premières composantes principales d'une ACP réalisée sur la base originale,
- la variable "Amount" correspondant au montant de la transaction,
- la variable "Time" qui décrit le laps de temps entre la transaction observée et la première transaction du jeu de données,
- la colonne "Class" indiquant si l'observation est une fraude ou si elle appartient à la classe des transactions normales.

4.2 Statistiques sur la base à disposition

Un grand déséquilibre est à noter sur les modalités de la variable "Class" en effet sur les 284807 transactions seulement 492 sont identifiées comme frauduleuses soit 0,17% des observations au globale, il faudra en tenir compte par la suite dans la modélisation.

4.2.1 Relation entre les variables explicatives et la variable cible

la figure 4.1 présente la distribution des différentes variables en fonction de la classe y [1] .

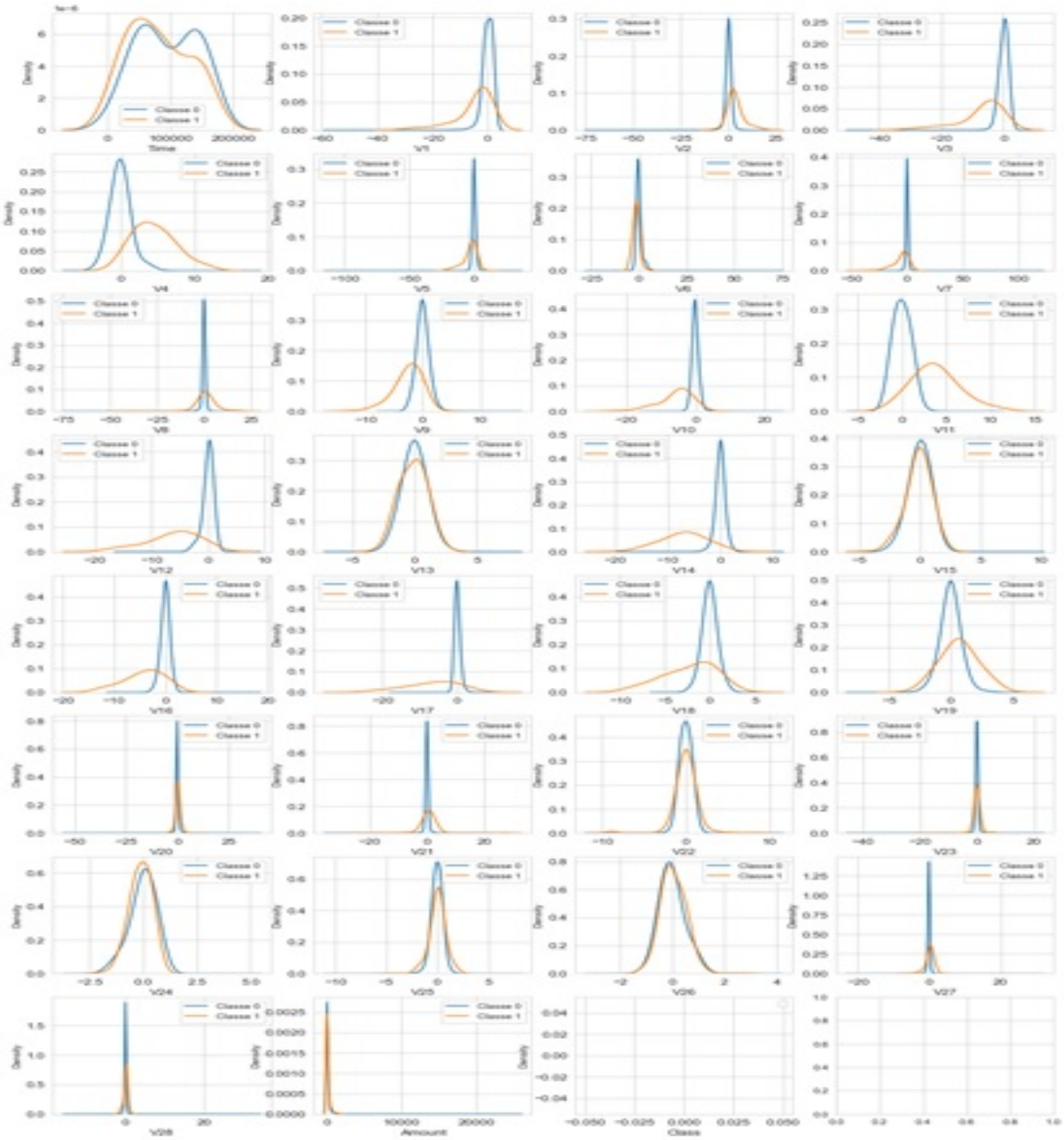


FIGURE 4.1 – Distribution des variables en fonction du type de classe

1. classe 0 observations normales ; classe 1 transactions frauduleuses

Les variables semblent avoir été standardisées dans la phase d'analyse en composante principale. En effet, les variables issues de l'ACP sont toutes symétriques et centrées en 0 pour la classe 0 (celle des observations non-frauduleuses).
Les variables Amount, V13, V15, V20, V22 à V26 et V28 semblent avoir des distributions similaires sur les 2 classes. Cependant, selon que l'on observe la classe 0 ou 1 la distribution des variables V4, V11, V12 et V18 semble différente. Cela pourrait indiquer que ces variables sont fortement liées à la fraude. Pour être rigoureux cette hypothèse sera testée via le test de Welch.
Le test t de Welch est une adaptation du test de Student permettant de comparer deux groupes lorsque leurs variances respectives sont différentes. Il s'agit donc de savoir si les moyennes des deux groupes sont significativement différentes d'un point de vue statistique. Soit A et B deux groupes à comparer alors la statistique de test T et le degré de liberté D sont données par les formules suivantes :

$$T = \frac{m_A - m_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

$$D = (\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B^2})/(\frac{S_A^4}{n_A^2\,(n_B - 1)} + (\frac{S_B^4}{n_B^2\,(n_B - 1)})$$

Avec,
m_A et m_B représentent la moyenne du groupe A et celui du groupe B, respectivement.
n_A et n_B représentent la taille du groupe A et celle du groupe B, respectivement.
S_A et S_B représentent l'écart-type du groupe A et celui du groupe B, respectivement.

Les résultats obtenus via l'outil ttest_ind du module scipy.stats de Python sont consignés dans le tableau 4.1 :

Variables	T	D
V4	-31.99	$3{,}18.10^{-139}$
V11	-29,30	$4{,}26.10^{-120}$
V12	-29,27	$3{,}05.10^{-113}$
V18	16,82	$7{,}93.10^{-52}$

TABLE 4.1 – Tableau Résultat du test de Welch

Pour réaliser ce test il est préférable d'avoir des classes équilibrées une transformation des données a donc été effectuée à ses fins.
L'hypothèse H0 d'égalité des moyennes entre les transactions frauduleuses et non-frauduleuses est donc rejetée, ces variables semblent donc bel et bien prendre des valeurs significativement différentes en fonction de la classe de la transaction.

4.2.2 Relation entre variables explicatives

La figure 4.2 représente la matrice de corrélation élaborée sur la base du coefficient de corrélation de Pearson. Elle permet de quantifier une quelconque relation linéaire entre les variables.

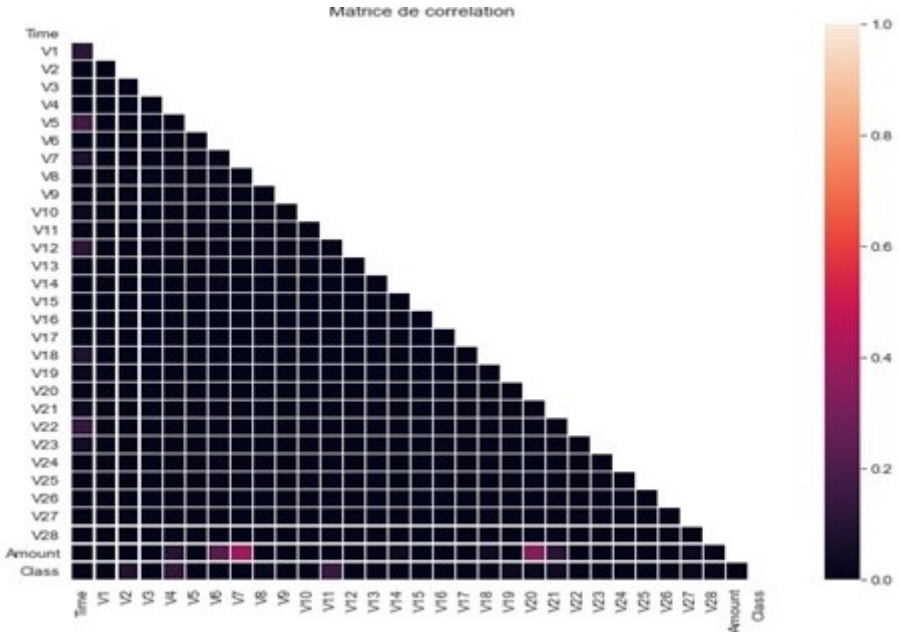


FIGURE 4.2 – Matrice de corrélation des variables

Il ne semble pas exister de relations entre les variables explicatives en effet aucune corrélation en valeur absolue n'est supérieur à 40%.

Chapitre 5

Modélisation

5.1 Apprentissage supervisé

Les algorithmes d'apprentissage supervisé extraient de la connaissance à partir de couples variables explicatives/variable à expliquer. Les données utilisées sont donc de la forme {Di = (Y_i,$X_{i,1}$,. . . , $X_{i,p}$) ; i = 1...n} où n et p représentent respectivement le nombre d'observation et le nombre de variables explicatives.

Le fonctionnement de ces modèles suppose une fonction de lien f entre les valeurs d'entrée X $\in R^{n\times p}$ et les valeurs de sortie Y$\in R^n$.Cette relation est caractérisée par les formules suivantes :

$$Y = f(X) + \epsilon$$

$$E[Y_i/X_{i,1},.....,X_{i,p}] = f(X_{i,1},.....X_{i,p})$$

La fonction f est inconnue à priori et représente l'information que X fournit sur Y. De plus, $\epsilon \in R^n$ représente le terme d'erreur aléatoire tel que $E(\epsilon) = 0$, et ϵ est indépendant de X. L'approche consiste à estimer la fonction f par une approximation $\hat{f}$. La prédiction $\hat{Y}$ aura la forme :

$$\hat{Y} = \hat{f}(X)$$

La fonction $\hat{f}$ est donc construite de façon à se rapprocher de f sans apprendre le bruit aléatoire ϵ.

5.2 Mesure de la qualité des modèles

5.2.1 Dilemme biais-variance

La fonction $\hat{f}$ est construit de manière à minimiser les écarts avec la vraie fonction de lien f. Cet écart se traduit mathématiquement par :

$$\begin{aligned} E[(f(X) - \hat{f}(X))^2)] &= E[(f(X) - E[\hat{f}(X)]) + (E[\hat{f}(X)] - \hat{f}(X))]^2 \\ &= E[(f(X) - E[\hat{f}(X)])^2] + E[(E[(\hat{f}(X)] - \hat{f}(X))^2]+ \\ &2 \times E[(f(X) - E[\hat{f}(X)])(E[(\hat{f}(X)] - \hat{f}(X))] \end{aligned}$$

En considérant les relations suivantes :

Biais$(\hat{f}(X)) = (f(X) - \hat{f}(X))$

Var$(\hat{f}(X)) = E[(\hat{f}(X) - E[\hat{f}(X)])^2] = E[\hat{f}(X)^2] - E[(\hat{f}(X)]^2$

$E[E[\hat{f}(X)]] = E[\hat{f}(X)]$

Alors l'écart se définit par la formule suivante :

$$E[(f(X) - \hat{f}(X))^2)] = \sigma^2 + Var(\hat{f}(X)) + Biais^2(\hat{f}(X))$$

Avec,

σ^2 correspondant à l'erreur incompressible qui permet de capturer la mesure dans laquelle l'estimateur varie en fonction de la base d'entraînement .

En d'autres termes un bon modèle de machine learning est celui ayant un fort pouvoir explicatif (faible biais) et une facilité à être généralisé (faible variance).

La figure 5.1 issue de l'article [16] de *Scott Fortman-Roe* illustre bien ce fait.

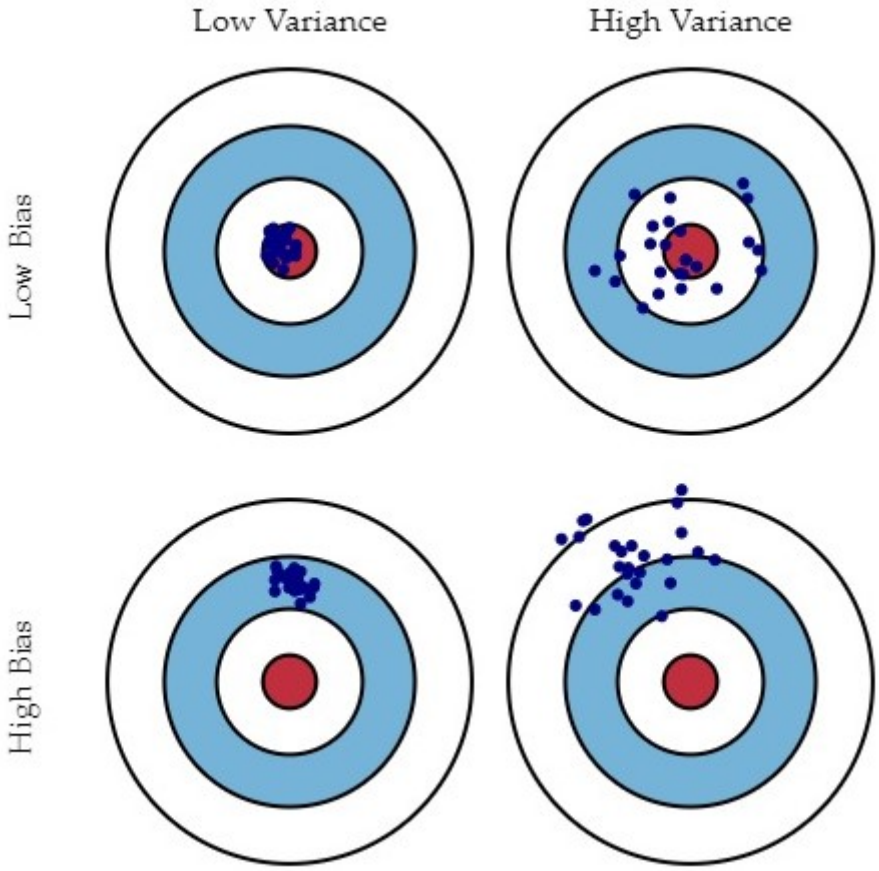


FIGURE 5.1 – Illustration du compromis biais-variance

En rouge la cible correspond aux vraies valeurs. Le modèle le plus performant est bien celui avec une faible variance et un biais faible.

5.2.2 Notions de sur-apprentissage

Un bon modèle de machine learning, est un modèle qui réalise de bonnes performances à la fois sur la base qui a servi pendant sa phase d'apprentissage mais également sur de nouvelles données, il s'agit là de la définition du concept clé de généralisation.
On parle de sur-apprentissage lorsque le modèle construit est trop proche des données d'apprentissage mais a du mal à être performant sur de nouvelles données. Un tel modèle va capter les bruits et les spécificités des données qu'on lui a fournies et aura donc du mal à se généraliser sur de nouvelles données. La figure 5.2 ci-dessous illustre ce phénomène.

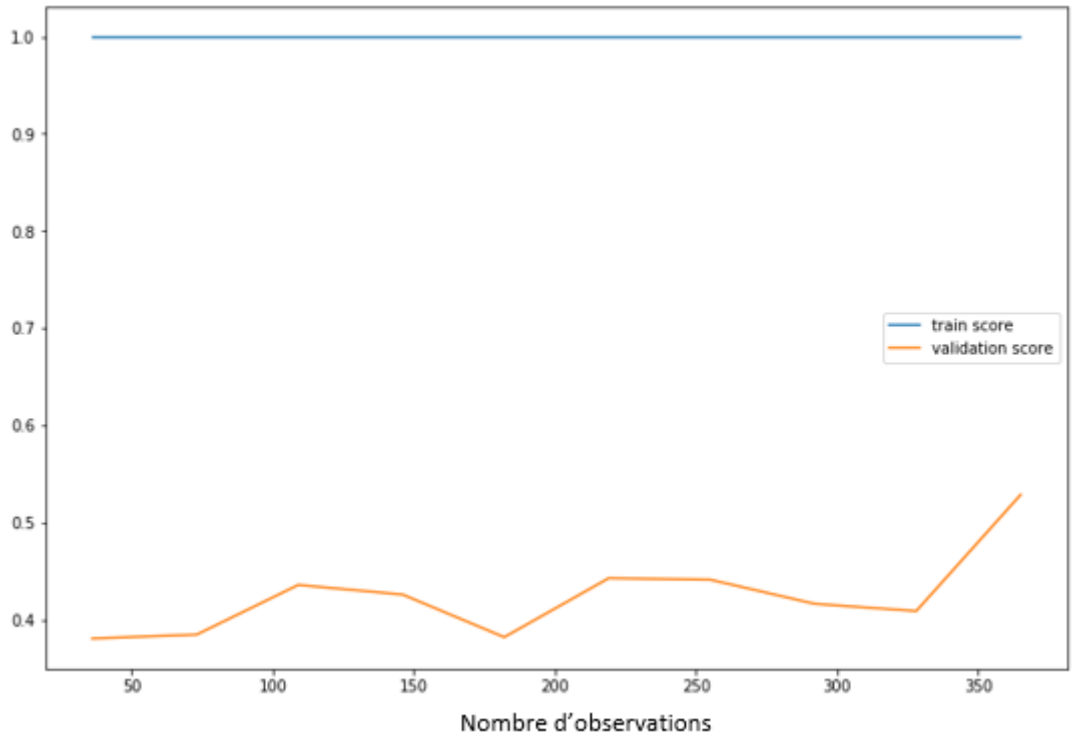


FIGURE 5.2 – Illustration d'un modèle en sur-apprentissage

Sur ce graphique est représenté le score F1 sur la base d'apprentissage en bleu et ce même score sur la base de test, il sera présenté dans la suite de ce mémoire à quoi correspond le score F1.

La notion de sur-apprentissage peut s'apparenter au dilemme biais-variance. L'explication est imagée sur le graphique 5.3 suivant :

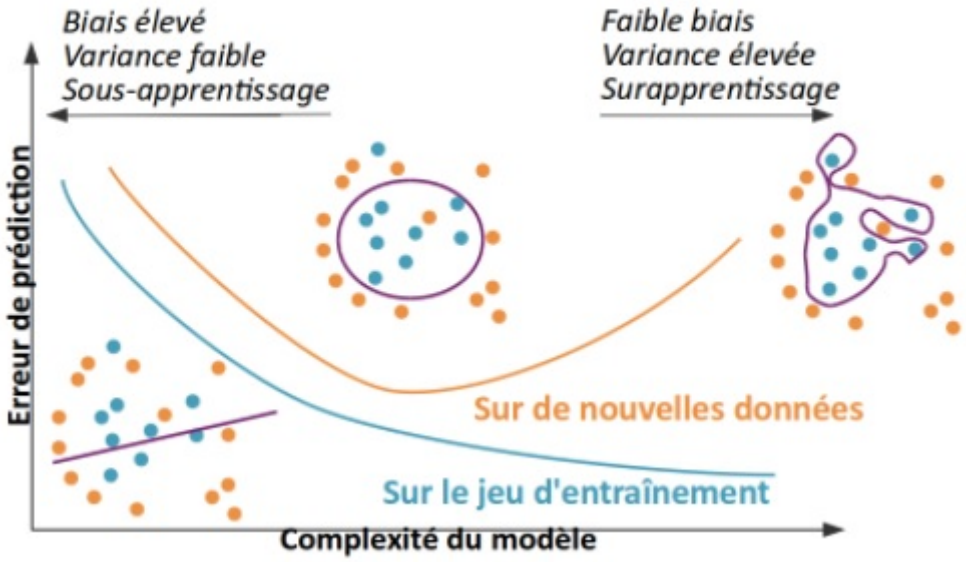


FIGURE 5.3 – Lien entre complexité et erreur de prédiction, source [14]

Lorsque la complexité du modèle augmente l'erreur diminue sur le jeu d'apprentissage, mais le modèle n'essaie pas dans ce cas d'apprendre le phénomène étudié mais plutôt de se rapprocher au mieux de la base d'entraînement ce qui aura pour conséquence d'augmenter les erreurs de prédictions sur une nouvelles bases. Les modèles en sur-apprentissage ont donc les caractéristiques suivantes :

— biais faible,
— variance élevée.

Le meilleur modèle (au milieu) est un compromis entre une complexité suffisante pour capturer la nature du problème étudiée (réduction du biais) et à la fois une complexité simple pour éviter le sur-apprentissage.

5.2.3 Validation croisée

Les critères de qualification d'un bon modèle d'apprentissage supervisée ont été présentés dans la section précédente. Pour évaluer la qualité d'un modèle, la procédure serait donc de constituer à partir du jeu de données à dispositions deux bases une base d'apprentissage et une base de test.
Cette vision sert pour comparer deux algorithmes différents mais dans la situation où on recherche à améliorer un modèle en optimisant ses hyperparamètres on ne peut l'utiliser car optimiser (régler) le modèle en évaluant ses performances sur la base de Test reviendrait à corrompre cette dernière.
Une solution serait de procéder au découpage suivant :

FIGURE 5.4 – Découpage du jeu de données

Le modèle serait entrainé sur la base d'apprentissage (Train set), calibré sur la base de validation (Val set) puis tester sur la base de Test (Test set).
Cette approche présente des limites en effet en fonction du découpage du jeu de données le calibrage des paramètres peut s'avérer différent.
Face à cette situation, la solution serait d'utiliser la validation croisée. Elle consiste à entraîner puis calibrer le modèle sur plusieurs découpes possibles de la base d'apprentissage.

FIGURE 5.5 – Cross validation avec un nombre de split=5

La configuration choisie au final pour le modèle sera celle qui en moyenne réalise de meilleures performances en fonction du critère choisit pour l'évaluation (score F1 par exemple).

5.2.4 Mesure de performance

La mesure de la performance dépend du problème auquel on est exposé. Dans le cadre d'un problème de régression par exemple un indicateur assez souvent utilisé est la MSE (Mean Squared Error) ou « EQM » (Erreur Quadratique Moyenne) définit par :

$$MSE = \frac{1}{m} \times \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Dans une approche liée à la classification binaire comme c'est le cas dans ce projet cet indicateur n'est pas adapté. Les indicateurs utilisés dans ce genre de problème dépendent de la nature des sorties de l'algorithme.

Cas des algorithmes qui retournent des valeurs binaires

La matrice de confusion ci-dessous permet d'évaluer les différents types d'erreurs.

	Classes prédites	
Classes réelles	Normales	Fraudes
Normales	Vrai Négatifs	Faux Positifs
Fraudes	Faux Négatifs	Vrai Positifs

FIGURE 5.6 – Matrice de confusion

De cette matrice plusieurs métriques de performance peuvent être déduites, on peut en citer :

— le nombre d'erreurs : il correspond à la somme entre les faux positifs et les faux négatifs,

— le taux de succès global ou Accuracy :

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

— la spécificité qui indique la capacité à détecter correctement le taux de vrai négatif

$$TVN = \frac{VN}{VN + FN}$$

— le recall ou sensibilité : il s'agit du taux de vrais positifs qui correspond ici à la proportion d'anomalies détectées. Ce taux est défini par :

$$TVP = \frac{VP}{VP + FN}$$

— la précision : cette métrique a pour objectif de minimiser le taux de faux positif.

$$Precision = \frac{VP}{VP + FP}$$

En fonction de la problématique le seuil voulu pour la précision ou le recall peut être différent. Prenons le cas d'un algorithme censé détecter si un malade est atteint de la covid-19 ou non. Maximiser la précision revient à minimiser le taux de faux positifs en d'autres termes des personnes déclarées porteuse du virus mais qui ne le sont pas.

Le recall quant à lui cherchera à minimiser le taux de faux négatifs c'est-à-dire de personnes déclarées saines alors qu'elles sont porteuses du virus.

Le risque associé aux faux négatifs dans ce cas de figure semble clairement plus grand que celui des faux positifs. L'attention portée sur le recall et la précision sera donc différente dans ce cas de figure.

Ces deux mesures de performances présentent des limites. Imaginons un cas de figure où on observe majoritairement des données d'une classe (classe 1 par exemple) pendant la phase d'apprentissage.

Le modèle prédira donc quasiment que toutes les données sont positives ce qui aura pour

conséquence d'avoir un recall égal ou très proche de 1 mais la précision dans ce cas sera très mauvaise. Inversement un modèle prédisant rarement des cas positifs aura une précision égale ou proche de 1 et un recall faible.
Pour pallier à ces limites, il existe une mesure qui est un compromis entre la précision et le recall , il s'agit du score F1 qui est la moyenne harmonique de ces deux critères.

Le score F1 est défini par :

$$F1 = 2 * \frac{Recall * Precision}{Precision + Recall}$$
$$F1 = 2 * \frac{VP}{2VP + FP + FN}$$

Cas des algorithmes retournant des scores

Dans le cas où l'algorithme essaie d'estimer la probabilité d'appartenir à une classe il est d'usage d'observer la courbe ROC (Receiver Operator Characteristic) définie ci-dessous :

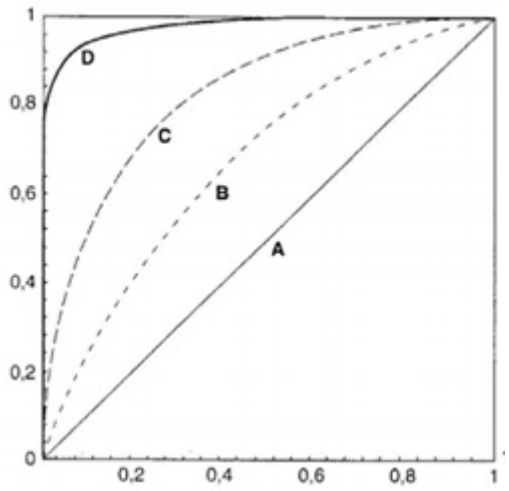


FIGURE 5.7 – Courbe ROC de différents modèles

En abscisse est observé 1-spécificité c'est-à-dire $\frac{FP}{FP+VN}$ qui correspond au taux de faux positifs en d'autres termes de fausses alertes qu'on cherche à minimiser (valeur proche de 0).
En ordonnée est représenté le recall qu'on cherche à maximiser.
La modèle parfait est donc celui dont la courbe se trouve dans le coin supérieur gauche en d'autres termes celui dont l'aire sous la courbe (AUROC[1]) est la plus élevée.

1. Area Under the ROC Curve

5.3 Choix des algorithmes

5.3.1 Comment choisir un bon modèle ?

Le choix de l'algorithme le mieux adapté à une problématique est fortement lié au type de données. La figure 5.8 donne des indications sur le choix à adopter parmi la large bibliothèque des modèles d'apprentissage.

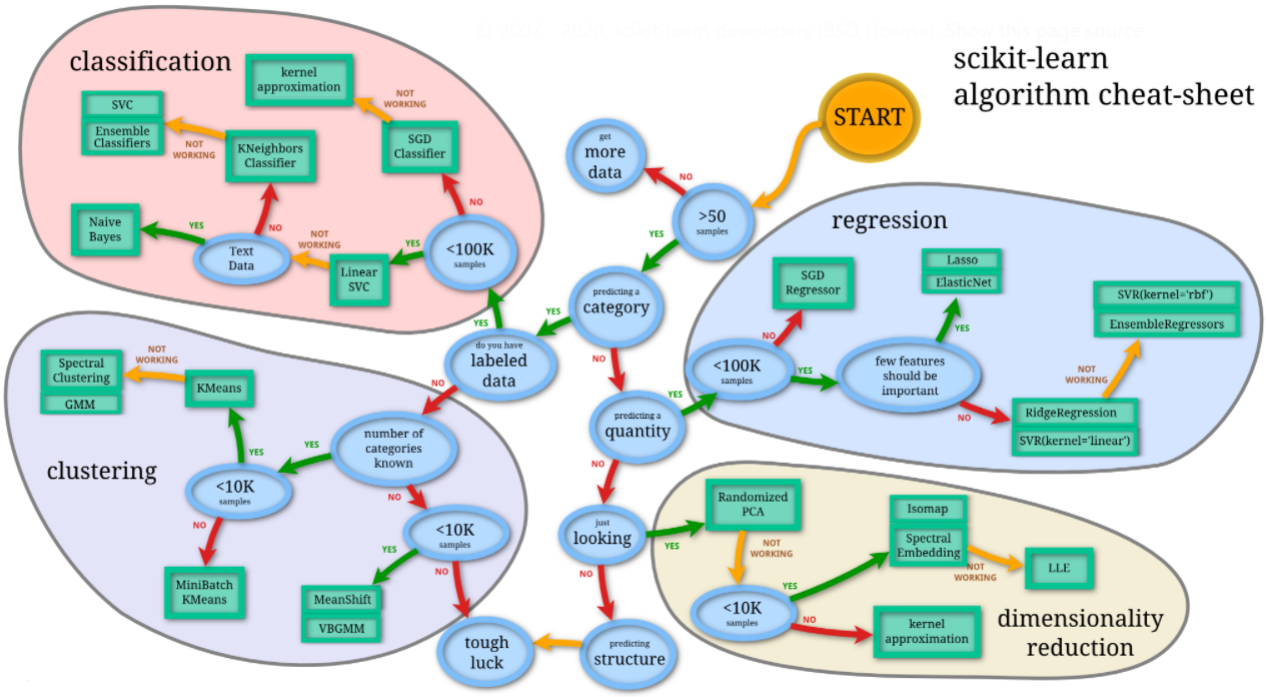


FIGURE 5.8 – Carte des algorithmes de scikit-learn, source [18]

Dans la pratique, il est difficile de définir une règle assurant la performance d'un modèle sur un type de données cela est dû à la l'infinité de possibilité de données. L'idée est donc de raisonner de façon pragmatique en testant une série de modèles et en conservant ceux réalisant les meilleures performances.

Dans le cadre de ce mémoire , les 5 modèles les plus performants seront présentés dans la suite il s'agit de 3 modèles d'ensemble (Random Forest, XGBOOST, STACKING) et deux modèles d'apprentissage non supervisé (Isolation Forest, Auto-encodeur).

5.3.2 Présentation théorique des modèles

5.3.2.1 Les Méthodes d'Ensemble Learning

« La majorité, dont chaque membre pris à part n'est pas un homme remarquable, est cependant au-dessus des hommes supérieurs »
Aristote-La politique

Les 3 méthodes retenues font partie des techniques d'Ensemble Learning qui répondent au concept du « Wisdom of the crowd » qui consiste à considérer qu'une foule d'individus est collectivement plus à même de résoudre un problème qu'un expert. Ces techniques consistent à entraîner plusieurs modèles de machine learning pour considérer l'ensemble de leurs prédictions. Pour que ces méthodes d'agrégations soient efficaces 3 conditions doivent être réunies par les modèles agrégés :

— un nombre assez suffisant,
— chaque modèle doit apporter un minimum d'informations,
— ces modèles doivent être diversifiés.

Il existe trois grandes techniques ensemblistes à savoir le bagging, le boosting et le stacking. Dans le cadre de ce mémoire il sera utilisé un modèle pour chacun de ces 3 ensembles.

Le Bagging

L'idée du bagging est d'entraîner à plusieurs reprises un algorithme (par exemple un arbre de classification) sur une portion aléatoire du jeu de données. L'échantillonnage des observations est réalisé par bootstrapping qui consiste en un tirage aléatoire avec remise des données. Le résultat final de la prédiction est obtenu par vote majoritaire dans le cas d'une classification ou par la moyenne dans le cas où la variable d'intérêt est continue.
Dans le cas où la variable cible est continue, nous obtenons avec le bagging une suite $(\phi_K)_{K=1.....B}$ de B arbres et l'arbre final retenu est défini par $\phi = \frac{1}{B}\sum_{K=1}^{B}\phi_K$. Le bagging permet de réduire la variance de l'estimation. En effet supposons que les arbres ϕ_K sont deux à deux corrélés par ρ et qu'ils ont une variance de σ^2 alors :

$$
\begin{aligned}
V[\phi] &= V[\frac{1}{B}\sum_{K=1}^{B}\phi_K] \\
&= \frac{1}{B^2}V[\sum_{K=1}^{B}\phi_K] \\
&= \frac{1}{B^2}\sum_{K=1}^{B}(\sigma^2 + \sum_{K'=1,K'\neq K}^{B}\rho\sigma^2) \\
&= \frac{\sigma^2}{B^2}\sum_{K=1}^{B}(1+(B-1)\rho) \\
&= \frac{\sigma^2}{B}(1+(B-1)\rho) \\
&= \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{B}
\end{aligned}
$$

La variance décroit en fonction du nombre d'itérations.
La figure suivante présente la structure de mise en place d'un algorithme de type bagging

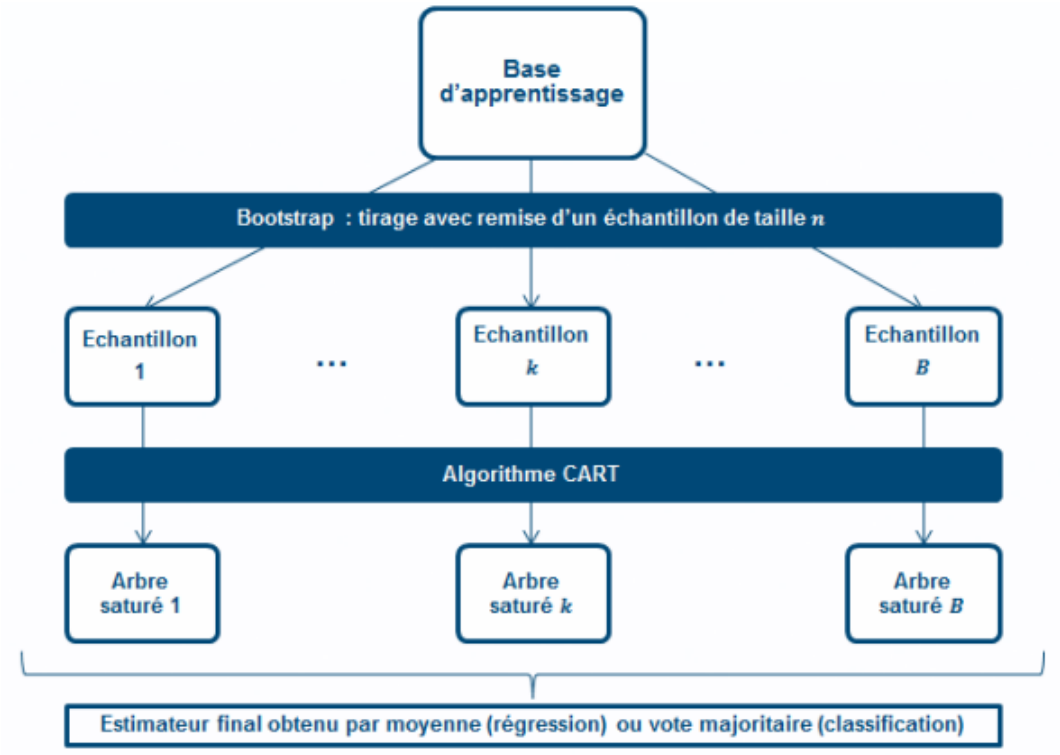


FIGURE 5.9 – Algorithme de bagging, source [1]

Une telle approche présente des limites dans le cas où les différents arbres construits sont dépendants. La méthode Random Forest permet de répondre à cette problématique.

Random Forest

Cet algorithme permet d'introduire de l'aléa dans la construction des différents arbres bootstrapés comme présenté dans la figure 5.10 ci-dessous :

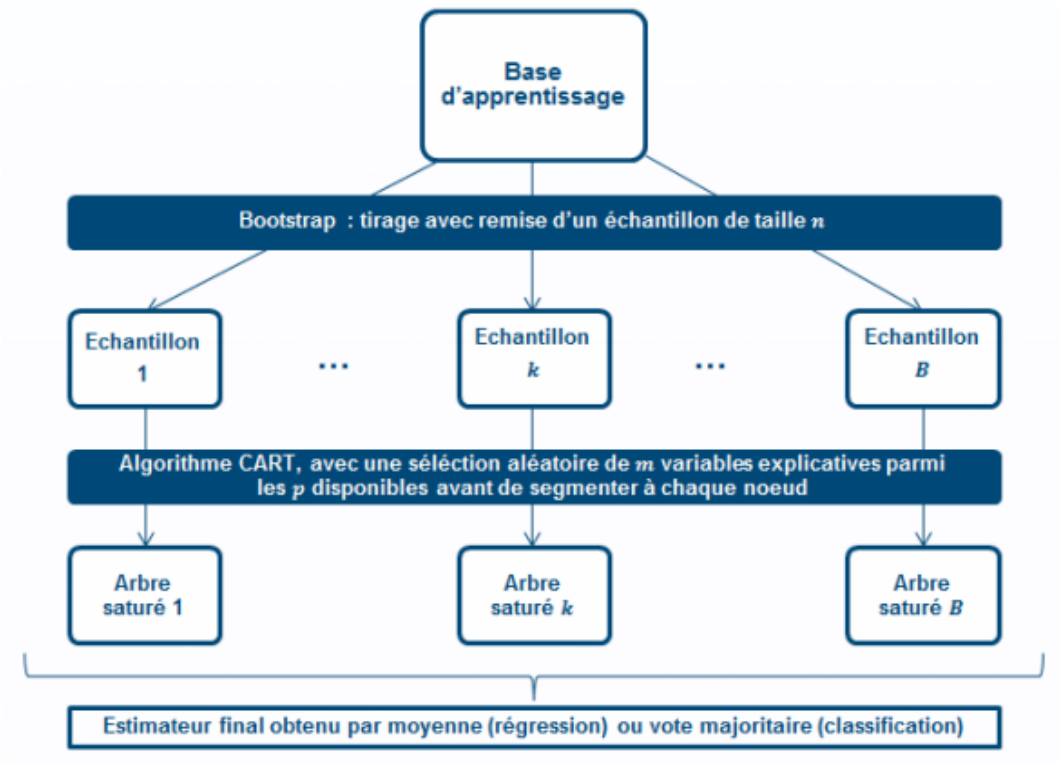


FIGURE 5.10 – Présentation de l'algorithme Random Forest, source [1]

Comme présenté précédemment la variance de l'estimateur final est définie par :

$$V(\phi) = \varphi\sigma^2 + \frac{(1-\rho)\sigma^2}{B}$$

Cette variance tend vers 0 si les arbres sont indépendants ie ρ tend vers 0 et B assez grand.

Random Forest en pratique

L'algorithme utilisé est RandomForestClasssifier du module sklearn.ensemble. De nombreux paramètres disponibles dans la documentation de scikit-learn peuvent être utilisés. Dans le cadre de ce mémoire ceux utilisés sont :

— n_estimators : le nombre d'arbres à entraîner qui est par défaut égale à 100,
— min_samples_split : le nombre minimal d'observations requis pour scinder un nœud qui est par défaut égale à 2,

— max_features : le nombre de variables considéré lors de la construction de l'arbre par défaut égale à $\sqrt{p}$ pour p variables explicatives,
— criterion : la fonction de mesure de la qualité de chaque split de l'arbre . Par défaut il s'agit du critère de Gini.

L'indice de Gini au nœud t se définit par :

$$i(t) = \sum_{k=1}^{K} P_{t,k}(1 - P_{t,k}) = 1 - \sum_{k=1}^{k} P_{t,k}^2$$

Avec ,
$P_{t,k} = \frac{n_{t,k}}{n_t}$ la proportion de la classe k dans le nœud t.

Le Boosting

Contrairement aux algorithmes de type "Bagging", l'idée ici est de construire plusieurs modèles simple (biais élevé/variance faible) en série, chaque modèle est construit de manière à corriger les erreurs du modèle précédent. On obtient à la fin un ensemble de modèles complémentaires dans lequel les faiblesses des uns sont compensées par les forces des autres.

XGBOOST

Le modèle XGBoost (Extreme Gradient Boosting) est une variante du Gradient Boosting , les principes énoncés ici sont tirés de la version évolutive [9].
Soit Θ l'espace des paramètres la fonction objective noté Φ à optimiser est définie de la façon suivante :

$$\Phi(\Theta) = L(\Theta) + \Omega(\Theta)$$

Avec L la fonction de perte issue de l'apprentissage et Ω un terme de régularisation.
La fonction L est un indicateur de la mesure de performance des prédictions du modèle.
Elle a plusieurs formes selon le phénomène étudié.
Dans un problème de régression il s'agit de l'erreur quadratique moyenne

$$L(\Theta) = \sum_i (y_i - \hat{y}_i)^2$$

Dans le cadre d'une classification L a la forme :

$$L(\Theta) = \sum_i [y_i \ln(1 + e^{-\hat{y}_i} + (1 - y_i) \ln(1 + e^{\hat{y}_i}]$$

Le terme de régularisation Ω quant à lui permet de contrôler la complexité du modèle. En effet comme énoncé plus haut l'algorithme est construit avec un modèle simple qui est en sous-apprentissage au départ (variance faible/biais élevé) chaque étape de l'algorithme consiste donc

à se rapprocher des données le terme Ω est une pénalité qui permet d'éviter le sur-apprentissage. La fonction objectif Φ est un compromis entre le biais et la variance, la fonction de coût L diminue lorsque la complexité Ω augmente.

Dans le cadre du XGBoost les différents modèles agrégés sont des arbres, la prédiction finale $\hat{y}_i$ s'écrit mathématiquement :

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F$$

Avec,
K : le nombre d'arbres.
f_k :le modèle construit à l'étape k (arbre k).
F : l'ensemble des modèles (arbres).

La fonction objective devient donc :

$$\Phi(\Theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Soit $\hat{y}_i{}^{(t)}$ la valeur de la prédiction à l'étape t, cette valeur est obtenue de façon itérative :

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

La fonction objective à l'étape t peut s'écrire donc :

$$\begin{aligned}
\Phi^t(\Theta) &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constante
\end{aligned}$$

En choisissant l'erreur quadratique moyenne comme fonction de perte on a :

$$\begin{aligned}
\Phi^t(\Theta) &= \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + constante
\end{aligned}$$

En pratique, on passe par un développement de Taylor à l'ordre 2 afin d'approcher la fonction objective :

$$\Phi^t(\Theta) \approx \sum_{i=1}^{n}[l(y_i,\hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t) + constante$$

Avec :

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)})$$
$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i,\hat{y}_i^{(t-1)})$$

En enlevant les contantes on a la formulation :

$$\Phi^t(\Theta) = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t)$$

Cette fonction est donc celle qu'on cherche à optimiser à chaque nouvel arbre consitué.

Zoom sur la complexité du modèle

Redéfinissons f(x) :

$$f_t(x) = w_{q_x}, w \in R^T, q : R^d \longrightarrow 1,2...,T.$$

Avec
w :le score associé à chaque feuilles
q : la fonction d'assignation de chaque point à la feuille correspondante
T : le nombre de feuilles

La complexité de l'abre peut se rédefinir donc par :

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

La fonction objective de l'arbre t se redéfinit donc par :

$$\Phi^{(t)}(\Theta) \approx \sum_{i=1}^{n}[g_i w_{q(x_i)} + \frac{1}{2}h_i w^2_{q(x_i)}] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$
$$\approx \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in i_j} h_i + \lambda)w_j^2] + \gamma T$$

Où $I_j = \{$i tel que $q(x_i) = j$ en d'autres termes tel que x_i appartient à la feuille j$\}$ est l'ensemble des observations assignées à la feuille j.

En posant $G_j = \sum_{i \in I_j} g_i$ et $H_j = \sum_{i \in I_j} h_i$
Alors l'équation se simplifie on obtient :

$$\Phi^{(t)}(\Theta) = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T$$

Dans cette équation le terme $G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2$ est une forme quadratique, les w_j sont indépendants les uns des autres. Pour une structure d'arbre q donnée les solutions aux problèmes d'optimisations sont données par :

$$w_j^\star = -\frac{G_j}{H_j + \lambda}$$

$$\Phi^\star = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

$\Phi^\star$ permet de mesurer la performance de la structure de l'arbre q(x).

Dans l'idéal, il serait intéressant d'énumérer toutes les possibilités d'arbre et de choisir le meilleur en se basant sur la mesure de performance mise en évidence précédemment. Compte tenu de l'infinité des structures d'arbres possible, en pratique on essaie d'optimiser un niveau d'arbre à la fois. Pour ce faire à chaque ajout d'un nœud à une feuille, la feuille est divisée en deux et la mesure du gain de score est définie par :

$$Gain = \frac{1}{2}[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}] - \gamma$$

Avec,

$\frac{G_L^2}{H_L+\lambda}$, $\frac{G_R^2}{H_L+\lambda}$ mesure le score sur la partie gauche et droite.

$\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ le score sur la feuille d'origine.

Le terme λ est la régularisation suite à l'ajout de complexité (séparation), en d'autres termes si le gain (en terme de score) est inférieur à λ il ne vaut mieux procéder à la séparation.

Implémentation de l'algorithme

La fonction « XGBClassifier » du package « xgboost » permet d'implémenter l'algorithme xgboost. Les paramètres utilisés sont :

— n_estimators : nombre d'arbres entrainés,
— min_child_weight(=1 par défaut) : nombre minimum d'observations par feuille par défaut,

— max_depth(=6 par défaut) : profondeur maximale d'un arbre, plus la valeur est grande plus on augmente la complexité du modèle et plus on s'expose au risque de sur-apprentissage,
— scale_pos_weight (=1 par défaut) : ratio entre le nombre d'observations de la classe négative et le nombre d'observations de la classe positive (fraude dans notre cas). Ce terme est très important en présence de classes déséquilibrées comme c'est le cas de cette étude.

STACKING

L'idée du stacking est d'entraîner un modèle de machine learning par-dessus les prédictions d'autres modèles. La figure décrit le fonctionnement de cet algorithme.

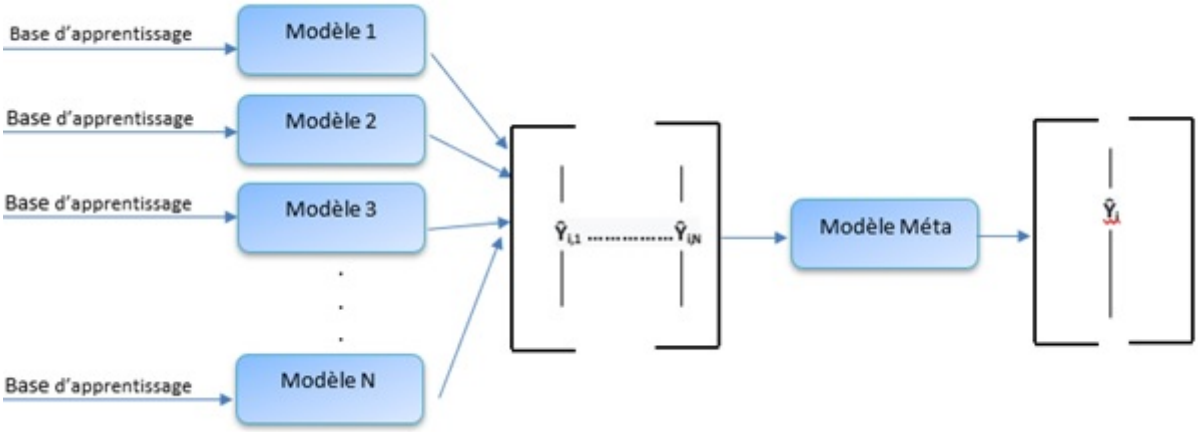


FIGURE 5.11 – Exemple d'un schema de stacking

On réalise dans un premier temps via différents modèles des prédictions sur la base d'apprentissage. Dans un second temps, au lieu de juste agréger ces prédictions un modèle méta sera entraîné pour prédire laquelle des prédictions des modèles est la bonne.
L'idée derrière est que le "Modèle Meta" puisse corriger les erreurs des modèles de la première étape. Une des conditions du bon fonctionnement de l'algorithme est que les modèles dans la phase 1 soient suffisamment différents.

Implémentation de la méthode Stacking

L'implémentation de cet algorithme est réalisé avec la classe StackingClassifier du module sklearn.ensemble. Les paramètres utilisés sont les suivants :

— estimators : liste des estimateurs utilisés en première étape,
— final_estimator(par défaut une régression logistique) : le modèle meta utilisé pour la prédiction finale.

Dans le cadre de l'étude, 2 modèles ont été implémentés en amont il s'agit du modèle XGBoost et Random Forest, le méta modèle choisit est la régression logistique. Ce choix n'est pas anodin en effet pour que cet algorithme soit le plus performant il est conseillé de choisir des modèles fonctionnant de façon différente.

Auto-encoder

Un Auto-encodeur est un réseau de neurones artificiels dont l'objectif est de reproduire les données qu'on lui fournit en entrée. La figure 5.12 schématise le processus de reconstruction de l'entrée.

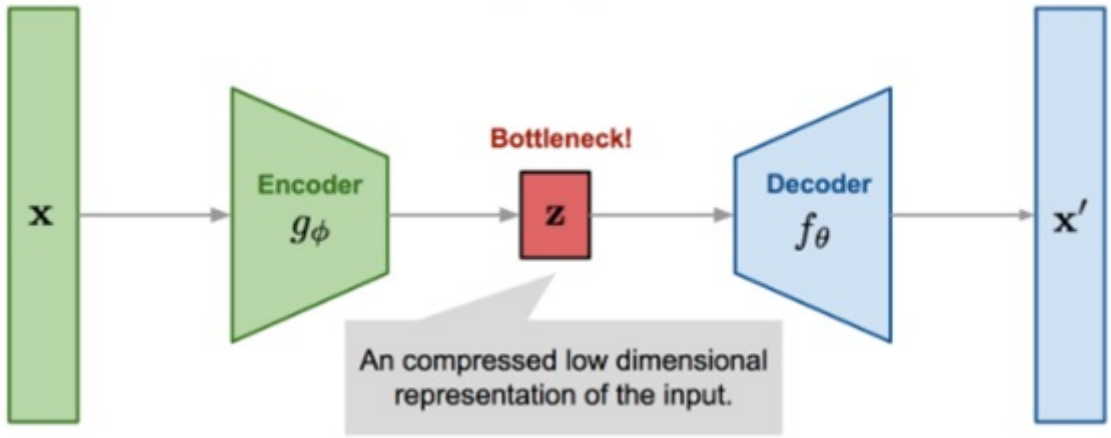


FIGURE 5.12 – Illustration de l'architecture d'un Autoencodeur, source [19]

X et X' représentent les données en entrées et en sortie, l'Auto-encodeur est constitué de plusieurs parties :

- un encodeur : il s'agit d'un réseau de neurones dont l'objectif est de compresser les données dans une dimension inférieur,
- une dimension latente : il s'agit de la représentation des données originales réduites. Cette représentation réduite de par sa construction sera concentrée uniquement sur les variables explicatives les plus importantes,
- un décodeur : ce réseau tente de reproduire les données en entrées à partir de l'espace des données réduites.

L'efficacité de cet algorithme va dépendre de certaines conditions parmi lesquelles on peut citer l'existence d'une corrélation entre les variables. En effet si toutes les variables sont fortement indépendantes il serait difficile de projeter les données sur un espace réduit sans perdre une quantité importante d'informations.
Soit,
g : la fonction d'encodage.

f : la fonction de décodage.
ϕ : l'espace des paramètres de l'encodeur.
θ : l'espace des paramètres du décodeur.
X : l'espace des données en entrées.
Z : la représentation de l'espace latent.
X' : les données en sorties.

Alors :

$$\phi.X \longrightarrow Z$$

$$\theta.Z \longrightarrow X'$$

De par ses relations on en déduit l'équation de la fonction coût L :

$$\begin{aligned} L(\phi,\theta) &= L'(x,x') \\ &= L'(x,f_\theta g_\phi(x)) \end{aligned}$$

Dans le cas d'une classification, la fonction L' correspond à la cross-entropy :

$$L'(x,f_\theta g_\phi(x)) = \frac{1}{m}\sum_{i=1}^{m}(-x_i \log f_\theta g_\phi(x_i) - (1-x_i)\log(1-f_\theta g_\phi(x_i))$$

Avec ,
m : le nombre de données en entrées
Dans le cas d'une régression la fonction coût s'écrit de la façon suivante :

$$L(\theta,\phi) = \frac{1}{m}\sum_{i=1}^{m}(x_i - f_\theta(g_\phi(x_i)))^2$$

La fonction dépend des paramètres ϕ ,θ qui sont obtenus par back-propagation.
Les applications des auto-encodeurs sont multiples on peut citer :

— le débruitage de données, en effet la phase de compression permet d'éliminer les informations superflues les données reconstruites à partir de cette base contiennent donc moins de bruit,
— la détection d'anomalies, ce thème sera abordé plus tard,
— la réduction de dimension est l'une des applications des auto-encodeurs. Ils présentent l'avantage de pouvoir appréhender des relations non-linéaires sur un jeu de données complexes contrairement à une analyse en composantes principales

L'auto-encodeur et l'ACP ont la même structure si l'on utilise des fonctions linéaires comme fonction d'activation.

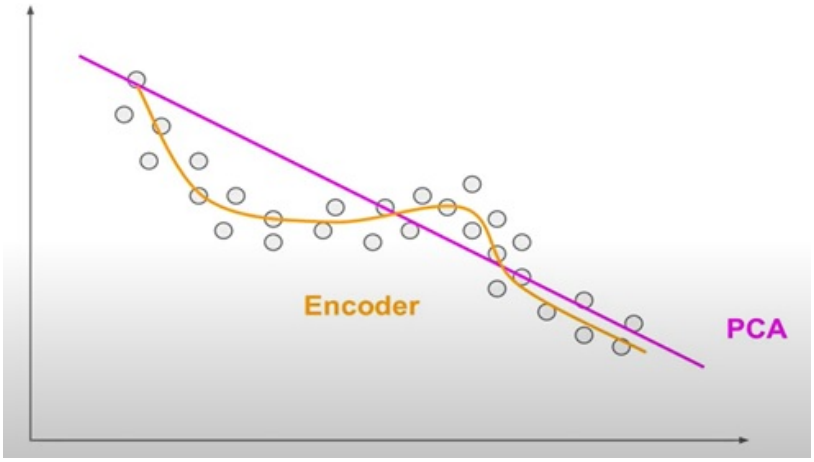


FIGURE 5.13 – Comparaison d'un ACP et d'un auto encodeur en 2 dimensions

L'Auto-encodeur variationnel

Il s'agit d'une variante des auto-encodeurs, ici l'espace latent n'est pas décrit par un vecteur fixe mais chacune de ses dimensions est représentée par une loi de probabilité.

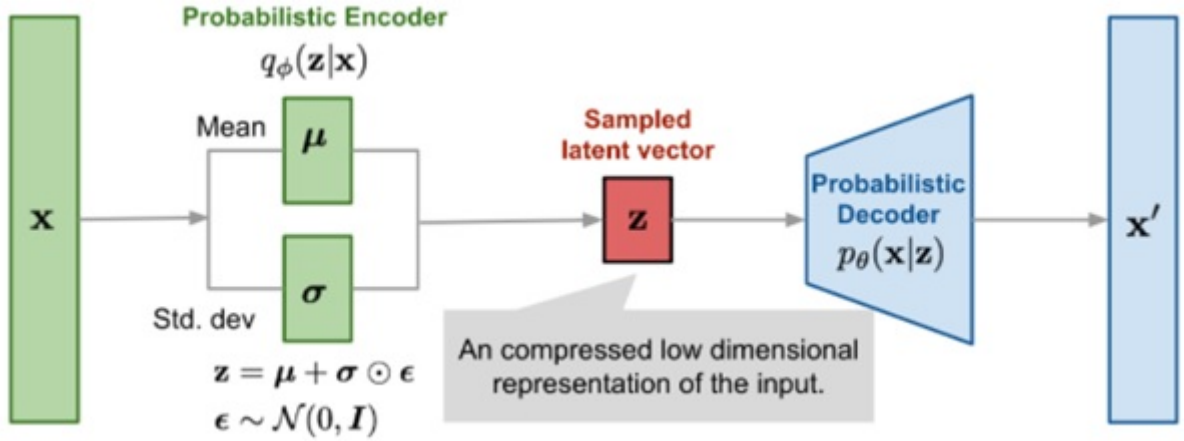


FIGURE 5.14 – Illustration d'un auto-encodeur variationnel sous l'hypothèse d'une gaussienne multivariée, source [19]

La figure décrit les étapes de la construction d'un auto-encodeur variationnel.
Le but de cet algorithme est de trouver une distribution de probabilité $q_\phi(z/x)$ des variables latentes, variables dont un échantillon sera choisi au hasard afin de générer de nouvelles observations x' via une probabilité $p_\theta(x/z)$. Les paramètres de ces distributions ϕ,θ sont obtenus également par back-propagation.
La probabilité $p_\theta(x/z)$ peut s'averer impossible à calculer.

Soit, X les données en entrée, Z l'espace des variables latente et $p(Z/X)$ définit par :

$$p(Z/X) = \frac{p(X/Z)p(Z)}{p(X)}$$

Avec,

$$P(X) = \int P(X|z)P(z)dz$$

En raison des intégrales multiples dont le nombre est lié à la dimension de Z l'intégrale peut s'avérer intraitable.
L'alternative consiste donc à approximer $p_\theta(x/z)$ par une distribution $q_\phi(x/z)$ via la méthode d'inférence variationnel. Cette méthode consiste à modéliser $p_\theta(x/z)$ en utilisant $q_\phi(x/z)$, avec $q_\phi(x/z)$ suivant une distribution assez simple comme une gaussienne par exemple.
L'objectif serait que ces distributions soit le plus proche possible. La proximité peut être mesurée via le critère de divergence de Kullback-Leibler qui permet de quantifier la distance entre deux distributions en mesurant la quantité d'information perdue si la distribution q est utilisé pour simuler p.
La divergence entre Q et P se définit de la façon suivante :

$$D_{KL}(Q||P) = E_{z\sim Q(z)} \log \frac{Q(z)}{P(z)}$$

Développons l'équation :

$$\begin{aligned}
D_{KL} q_\phi(z|x)||p_\theta(z|x)) &= \int q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z|x)} dz \\
&= \int q_\phi(z|x) \log \frac{q_\phi(z|x)p_\theta(x)}{p_\theta(z,x)} dz \\
&= \int q_\phi(z|x)(\log p_\theta(x) + \log \frac{q_\phi(z|x)}{p_\theta(z,x)})dz \\
&= \log p_\theta(x) + \int q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(z,x)})dz \\
&= \log p_\theta(x) + \int q_\phi(z|x) \log \frac{q_\phi(z|x)}{p_\theta(x|z)p_\theta(z)} dz \\
&= \log p_\theta(x) + E_{z\sim q_\phi(z|x)}[\log \frac{q_\phi(z|x)}{p_\theta(z)} - \log p_\theta(x|z)] \\
&= \log p_\theta(x) + D_{KL}(q_\phi(z|x)||p_\theta(z)) - E_{z\sim q_\phi(z|x)} \log p_\theta(x|z)]
\end{aligned}$$

On a,

$$D_{KL}(q_\phi(z|x)||p_\theta(z|x)) = \log p_\theta(x) + D_{KL}(q_\phi(z|x)||p_\theta(z)) - E_{z\sim q(z|x)} \log p_\theta(x|z)$$

$$\log p_\theta(x) - D_{KL}(q_\phi(z|x)||p_\theta(z|x)) = E_{z\sim q_\phi(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p_\theta(z))$$

Le terme de droite de l'équation est appelé fonction objective. Le premier terme correspond à la logvraisemeblance de reconstruction et le second terme mesure à quel point la distribution Q est proche de celle de P.
La fonction de coût est :

$$\begin{aligned} L(\theta,\phi) &= -\log p_\theta(x) + D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \\ &= -E_{z\sim q_\phi(z|x)} \log p_\theta(x|z) + D_{KL}(q_\phi(z|x)||p_\theta(z)) \end{aligned}$$

Minimiser donc cette fonction revient à maximiser la log-vraisemblance log $p_\theta(x)$.

Application à la détection d'anomalies

La figure ci-dessous décrit les étapes de la mise en place d'un auto-encoder dans la pratique.

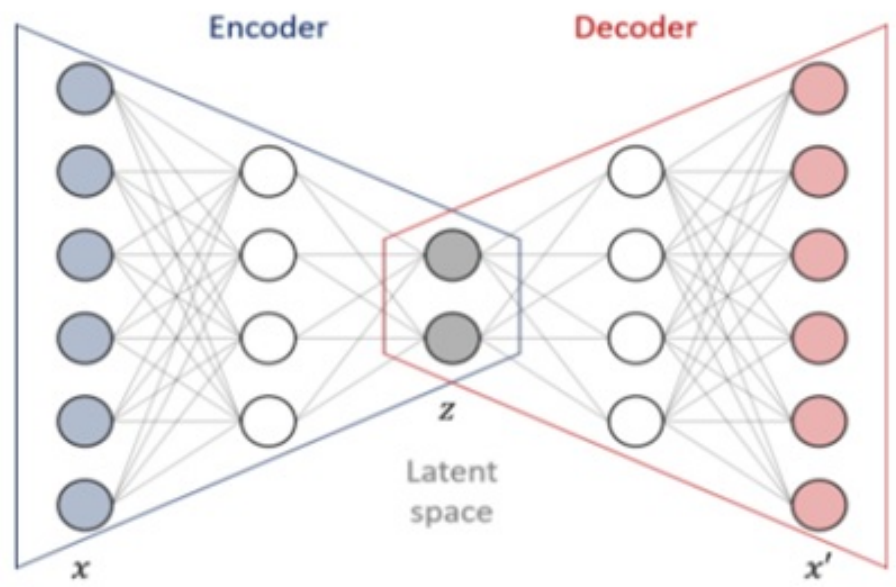


FIGURE 5.15 – Schéma d'un auto-encodeur

Le processus de décodage reflète le processus de codage dans le nombre de couches cachées et de neurones, afin de pouvoir réduire les dimensions des données les couches intérieurs sont plus petites que les couches extérieures.

En quoi cette structure permet-elle de détecter les anomalies ?

L'auto-encodeur par sa structure essaie de reproduire les données en entrées tout en cherchant à minimiser l'erreur de construction.
En pratique, la base d'apprentissage qui servira à construire l'auto-encodeur ne contiendra que

des transactions 'normales' afin que ce dernier ne puisse reproduire que ces dernières. Dans un second temps le modèle sera confronté à une base test contenant à la fois des transactions frauduleuses et normales. Les observations frauduleuses entrées en input auront une erreur de reconstruction supérieur aux observations normales.
L'auto-encodeur peut être utilisé dans un cadre non-supervisé mais cette approche ne sera pas étudiée dans le cadre de ce mémoire. En effet, compte tenu du fait que le jeu de données contient en majorité des observations 'normales', l'auto-encodeur dans sa quête de reproduction des données globales tout en minimisant les erreurs de reconstruction va donc essayer de reproduire les données dites 'normales'.
Lorsqu'une anomalie sera mis en input du modèle ce dernier aura donc du mal à reproduire, les données présentant une erreur de reconstruction plus élevées que les autres seront celles considérées comme anormales.

Structure du réseau de neurones

Dans le cadre de ce mémoire la structure choisie est un réseau avec 6 couches dont 4 intermédiaires définies de la façon suivante :

- une couche d'entrée et une couche de sortie de dimension égale au nombre de variables,
- couche 1(encodeur1) : le nombre de neurones égale à 14 avec la fonction d'activation tangente hyperbolique,
- couche 2(encodeur2) : couche permettant à la réduction de dimensions , le nombre de neurones est égale à 7 et la fonction d'activation choisie est RELU(REctified Linear Unit),
- couche 3(decodeur1) : le nombre de neurones est égale à 7 et la fonction d'activation est la tangente hyperbolique,
- couche 4(décodeur2) : le nombre de neurones est égale à celui de l'encodeur 1, la fonction d'activation est RELU.

5.4 Présentation des résultats

5.4.1 Apprentissage non supervisé

L'algorithme utilisé pour la détection des fraudes est l'Isolation Forest. Contrairement au cadre supervisé, l'apprentissage non-supervisé ne nécessite pas de diviser le jeu de données en données de test et données d'apprentissage. Il sera fourni donc à l'algorithme le dataset en entier puis les résultats des clusters seront confrontés aux vrais valeurs de la variable cible.

La figure 5.16 présente l'histogramme des scores d'anomalies, sur cette base la règle de décision a été de considérer comme anormales toutes les observations dont le score d'anomalies est supérieur à 65%.

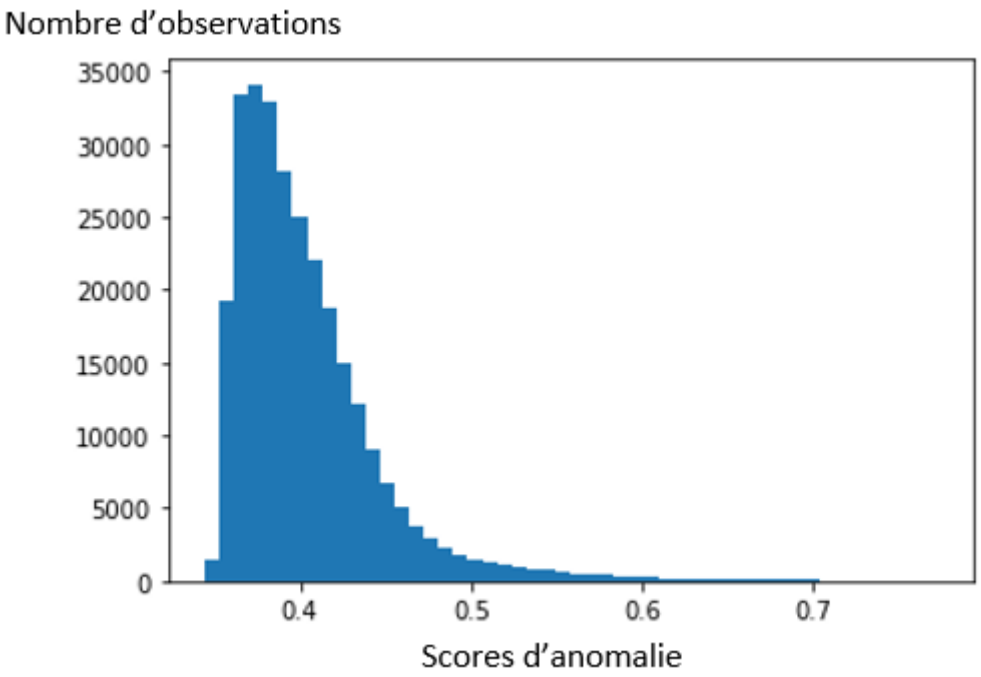


FIGURE 5.16 – Histogramme des scores d'anomalies

Le tableau suivant présente les résultats de l'implémentation de l'algorithme d'Isolation Forest.

	Prédictions	
Classes réelles	0	1
0	283889	426
1	321	171

Table 5.1 – Matrice de confusion de l'algorithme Isolation Forest

Cette matrice est très instructive sur la performance de l'algorithme.

Sur les 284210 transactions normales 283889 ont été jugées comme telles par l'algorithme soit un accuracy de 99% sur les transactions normales.

597 transactions ont été jugées comme anomalies par l'algorithme et parmi elles 171 ce sont avérées être des transactions frauduleuses parmi les 492 que compte la base soit un recall de 35 % (pourcentage de transactions frauduleuses détectées par l'algorithme).

Le taux de faux positifs observé est assez elevé 71% ce chiffre est à relativiser car l'algorithme est calibré pour détecter les anomalies et non des fraudes ces observations sont donc simplement des transactions différentes des autres mais qui ne sont pas automatiquement des fraudes, on peut penser à des retraits de montants assez élevés par exemple.
Le score F1 sur la base globale est de 31% ce qui semble très satisfaisant, car aucun paramétrage préalable n'a été nécessaire pour obtenir une telle performance.

5.4.2 Apprentissage supervisé

Avant d'appliquer les différents algorithmes présentés en 5.3.2 le jeu de données a été scindé en 2 bases, une base d'apprentissage (80%) et une base de test (20%). Cette séparation a été réalisée à l'aide de la fonction traintestsplit du module model_selection de scikitlearn. Cette fonction permet de scinder la base en faisant de la stratification c'est-à-dire en conservant la distribution relative des classes (ici frauduleuses ou non).

La figure 5.17 représente la matrice de confusion des 4 modèles entraînés :

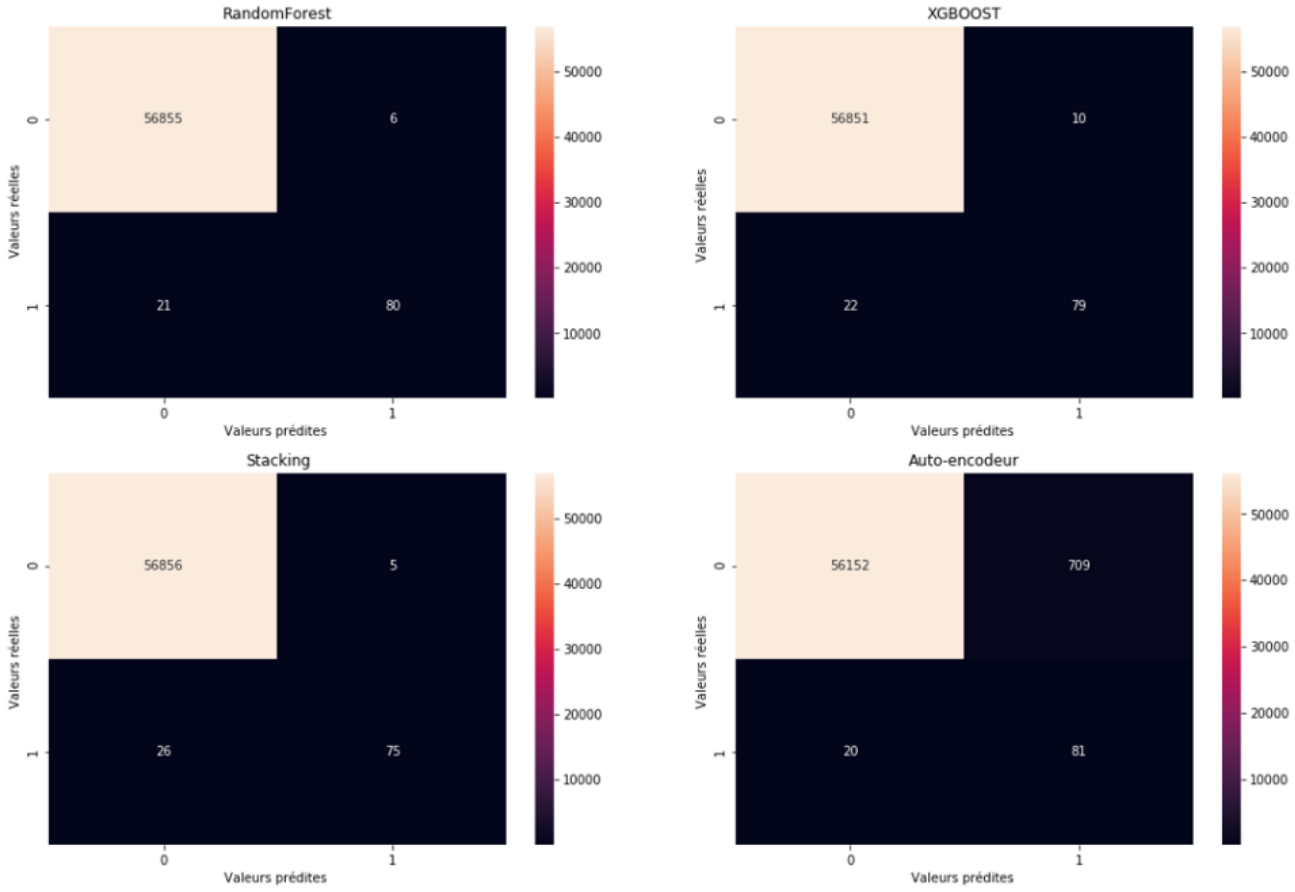


FIGURE 5.17 – Matrice de confusion des 4 modèles

Dans l'ensemble les modèles d'agrégation produisent peu d'erreurs. Ils arrivent à bien identifier les transactions selon leurs différentes classes respectives.

Le tableau ci-dessous résume la performance des différents modèles en fonction de plusieurs critères de performance.

	Précision	Recall	Score F1
Random Forest	93%	79%	86%
XGBOOST	89%	78%	83%
STACKING	94%	74%	83%
AUTO-ENCODEUR	10%	80%	18%

TABLE 5.2 – Tableau des performances

L'auto-encodeur semble exceller dans la détection d'observations frauduleuses car le taux de vrais positifs est de 80%, cependant le nombre de faux positifs est assez élevé, ce qui impact fortement la précision et donc le score F1 par la même occasion. Cette valeur s'explique principalement par le seuil de reconstruction choisit pour qualifier une transaction de frauduleuse. En effet une hausse de ce seuil entraîne une baisse de significative du nombre d'observations classées comme frauduleuses et par la même occasion le taux de faux positifs (hausse de la Précision), cependant cela augmente nettement le taux de faux négatifs (baisse du Recall).

A titre de comparaison, il est important de noter que sur la même base de test l'algorithme d'Isolation Forest obtient un score F1 moins important que sur la base globale de 28% .
Une explication à ce phénomène pourrait être que la séparation du jeu de données(Trainset/Testset) entraîne une perte d'informations, l'Isolation Forest agissant à "l'aveugle" a donc besoin d'un maximum possible de données pour arriver à comprendre au mieux leur structure afin d'établir sa cartographie.

5.5 Amélioration des performances : mixte entre l'apprentissage supervisé et la vision non-supervisée

5.5.1 Choix du modèle à améliorer et justification des approches considérées

La problématique de cette partie est de détecter les observations frauduleuses, la part de transactions frauduleuses réelles détectées (Recall) est plus élevée avec les algorithmes d'auto-encodeur et Random Forest. Cependant, le modèle Random Forest est l'algorithme ayant le score F1 le plus élevé.
Dans la suite de l'analyse, le choix sera donc d'améliorer les performances de cet algorithme.
Le recall de ce modèle est de 79% autrement dit, le modèle manque d'identifier 21% des transactions frauduleuses.
En observant de plus près ces transactions frauduleuses non détectées, on remarque que 5% d'entre elles sont détectées par l'algorithme d'Isolation Forest et 10% par l'auto-encodeur.
Explorer une approche mixte entre le modèle Random-Forest et les deux algorithmes d'Isolation Forest et Auto-encodeur pourrait s'avérer donc plus efficace pour améliorer la performance de ce modèle.

5.5.2 Méthode 1 : Rééquilibrage des données et utilisation du score d'anomalie de l'Isolation Forest

Le problème des jeux de données déséquilibrées

Pour rappel l'Accuracy se définit par :

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Les jeux de données déséquilibrés ont tendance à biaiser les modèles de prédictions. En effet les algorithmes de par leurs constructions cherchent à maximiser le taux de succès global. Dans une base avec un déséquilibre l'algorithme aura donc tendance à prédire majoritairement la classe en surnombre ce qui aura pour effet d'augmenter l'accuracy. Dans notre cas cela peut s'avérer fort handicapant car la classe majoritaire correspond aux transactions normales or le but de cette analyse est de détecter les transactions frauduleuses. Pour faire face à ce problème plusieurs solutions peuvent s'avérer efficace à savoir le sur-échantillonnage ou le sous-échantillonnage.
Le sous-échantillonnage consiste à diminuer le nombre d'échantillon de la classe majoritaire afin de rééquilibrer le jeu de données, cette technique possède un gros désavantage c'est la perte d'informations sur la classe majoritaire liée à la suppression d'observations.
Le sur-échantillonnage consiste à reproduire les observations appartenant à la classe minoritaire. Contrairement au sous échantillonnage on ne perd aucune information, cependant cette méthode expose les modèles de prédiction à l'overfitting. Certaines méthodes permettent de générer de nouveaux exemples artificiels proches d'un certain point de vue des observations de la classe minoritaire. C'est le cas de la méthode SMOTE.

SMOTE (Synthetic Minority Over-sampling Technique)

L'idée de cette méthode consiste à générer de nouveaux individus de la classe minoritaire. Ce processus s'effectue en traçant dans l'espace des variables explicatives un segment entre 2 observations rares et en plaçant aléatoirement sur ce segment de nouvelles observations. La figure 5.18 ci-dessous résume bien l'idée de cette méthode.

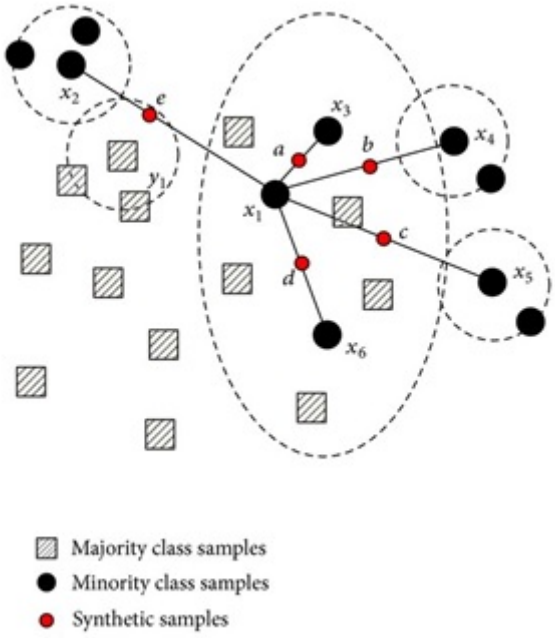


FIGURE 5.18 – Description du suréchantillonage via la méthode SMOTE, source [8]

La librairie Imbalanced-learn permet de mettre en place cette technique dans python.
Dans la pratique, l'algorithme de SMOTE fonctionne de la façon suivante :

Etapes de la mise en place de l'algorithme

1. Identifier les k plus proches voisins de l'observation minoritaire
2. Faire la différence entre les variables de l'observation minoritaire et un des k voisins (calcule de la distance entre les 2 observations)
3. Multiplier cette différence par un nombre tiré aléatoirement entre 0 et 1
4. Additionner le terme obtenu précédemment aux variables de l'observation

Comme mentionné dans l'article originale sur la méthode SMOTE il est conseillé pour avoir de meilleurs résultats de procéder à un sous-échantillonnage de la classe majoritaire.
La modélisation utilisée dans la suite sera effectuée sur une base d'apprentissage rééquilibrée via l'algorithme SMOTE.

Utilisation du score d'anomalie

Une fraude étant considérée comme une observation atypique, le score d'anomalie calculé par l'algorithme d'Isolation Forest représente une information car il donne une indication sur le caractère atypique d'une transaction.
L'idée développée est la suivante :

Etapes de la mise en place de l'algorithme

1. Calcul du score d'anomalie pour chaque observation de la base d'apprentissage
2. Rajout des scores d'anomalies comme variables explicatives.
3. Entrainement du modèle
4. Mesure des indicateurs de performances

Présentation des résultats

Le tableau suivant présente la différence de résultats entre le modèle Random Forest et le modèle 2 qui correspond à l'application de l'algorithme Random Forest sur la base d'apprentissage réequilibrée par l'algorithme SMOTE et enrichie par la variable explicative liée au score d'anomalies.

	Précision	Recall	Score F1
Random Forest	93%	79%	86%
Modèle 2	91%	85%	88%

Table 5.3 – Comparaison des résultats

Les modifications apportées entraîne une légère baisse de la Précision (hausse du taux de faux positifs) mais une hausse bien plus significative du Recall (baisse du taux de faux négatifs).
Le score F1 augmente de 2% pour atteindre 88%, cette approche semble donc assez efficace et adaptée à la problématique de détection de fraudes.

5.5.3 Méthode 2 : Prédiction via l'espace latent de l'Auto-Encodeur

Pour rappel le but de l'auto-encodeur est d'arriver à reproduire les données en entrée en passant par une phase de réduction de dimension.
La réduction de dimension effectuée a pour but de réduire la complexité superflue de l'input fourni afin que la reproduction réalisée ne s'intéresse qu'aux caractéristiques les plus importantes. Cette approche peut s'apparenter à la technique de sélection de variables les plus importantes.
Construire un algorithme de classification (implémentation du Random Forest) à partir de l'espace latent permettrait donc de faciliter l'apprentissage et lutter par la même occasion contre le fléau de la dimension. La technique utilisée est la suivante :

Etapes de la mise en place de l'algorithme

1. Récupération des données à partir de l'espace latent
2. Séparation des données en 2 bases (apprentissage et test)
3. Construction d'un algorithme de détection de fraudes sur la base d'apprentissage
4. Evaluation de la qualité de l'algorithme sur la base de test

Le tableau suivant sert d'élément de comparaison entre les résultats d'un auto-encodeur simple et celui proposé dans cette section .

	Précision	Recall	Score F1
Random Forest	93%	79%	86%
Autoencodeur simple	10%	80%	18%
Autoencodeur mixte	98%	60%	74%

Table 5.4 – Tableau de comparaison des modèles

Cette méthode augmente considérablement la précision, en effet le nombre de transactions

prédites comme frauduleuses sont beaucoup moins nombreuses mais ces dernières sont à 98% des cas de fraudes avérés (réduction importante du taux de faux positifs). Néanmoins l'algorithme manque d'identifier 40% des transactions frauduleuses, le taux de vrais positifs n'est égale qu'à 60%.
Comparativement à l'algorithme Random Forest, la baisse du taux de faux positifs est moins importante que la hausse du taux de faux négatifs, ce qui a pour effet une dégradation du score F1 à 74%.

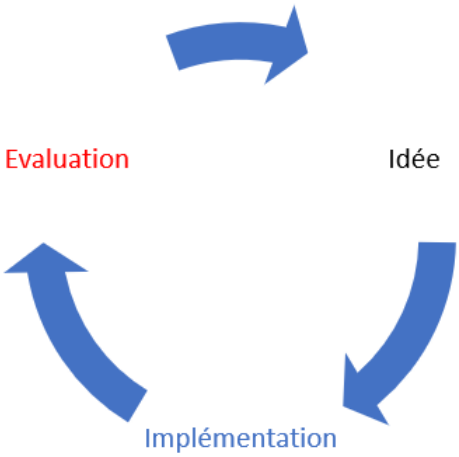
5.6 Synthèse des résultats

L'objectif de cette section a été de montrer l'efficacité des algorithmes de détection d'anomalies dans un cas pratique de détections de fraudes.
L'algorithme d'Isolation Forest dans sa quête de détection d'observations anormales à réussi à détecter 35% du total des transactions frauduleuses. Ses résultats sont assez convaincants en ce sens où l'algorithme n'est pas construit de manière à détecter directement des fraudes, mais juste les observations qui présentent une structure atypique. Sur l'ensemble des observations catégorisées comme anomalies 28% ce sont avérées être des fraudes . Cette faible précision est due à plusieurs facteurs :

— l'algorithme d'Isolation Forest met en évidence les anomalies, toute fraude est une anomalie mais toutes les anomalies ne sont pas des fraudes,
— les observations qualifiées de fraudes sont des fraudes avérées mais celles qualifiées comme non frauduleuses sont celles dont le caractère frauduleux n'a pas pu être prouvé(elles peuvent donc être frauduleuses),
— le nombre de faux positifs est fortement lié au choix du seuil choisi pour classer les transactions en fonction de leur score d'anomalie. En effet un choix de seuil inférieur aurait permis de déclarer plus d'observations comme frauduleuses et par le même mécanisme augmenté la qualité de prédiction des observations frauduleuses mais dans le même temps cela aurait diminué la qualité de prédiction des observations non-frauduleuses. Le tout est donc de trouver un bon compromis.

Sur la base des résultats obtenus par les différents algorithmes, le second objectif a été d'explorer les possibilités pour améliorer les performances des modèles en ayant une approche mixte entre apprentissage supervisé et non-supervisé.
La boucle itérative suivante résume bien la démarche adoptée dans cette section.

Chaque idée est mise en place puis évaluée afin d'en tester la pertinence.



En suivant cette voie deux approches intéressantes ont été implémentées :

— un modèle basé sur le critère quantitatif : il s'agit de l'implémentation du modèle Random Forest sur la base rééquilibrée et enrichie de la variable score d'anomalie de l'Isolation Forest. Ce modèle arrive à identifier 85% des transactions frauduleuses,

— Un modèle basé sur un critère qualitatif : il s'agit de l'implémentation du modèle Random Forest sur l'espace latent(base réduite), ce modèle a un fort taux de précision en effet 98% des observations déclarées comme frauduleuses par ce modèle le sont en réalité.

Plusieurs enseignements peuvent être tirés des performances de ces modèles. Sur la base considérée :

— la réduction de dimension permet d'améliorer la précision des modèles,

— l'information apportée par l'Isolation Forest permet de mieux détecter les transactions frauduleuses (en quantité).

5.7 Conclusion

L'objet de ce mémoire a été d'améliorer la qualité des données par la détection d'anomalies à travers des méthodes d'apprentissage non-supervisé notamment l'Isolation Forest un algorithme permettant de mettre en évidence au sein d'une base de données les observations dont les caractéristiques sont assez différentes (éloignées) des autres.

L'avantage de l'apprentissage non-supervisé est la facilité de sa mise en oeuvre car cette méthode ne présuppose aucun étiquetage préalable des données d'apprentissage. Néanmoins, le talon d'Achille de cette vision est l'impossibilité de back-tester les résultats. Pour cette raison, la dernière partie de ce mémoire a été consacrée à la détection de fraudes via l'approche non-supervisée sur une base étiquetée. Un tel choix est motivé par le faite qu'une fraude par nature est une observation qu'on ne pourrait pas qualifier de "normale". Les résultats observés sont très satisfaisants, l'algorithme d'Isolation Forest réussi à identifier plus d'une fraude sur 3 (35%) sans le moindre paramètrage préalable. Les travaux réalisés ont permis donc de montrer qu'il existe bel et bien un lien entre fraudes et anomalies, ils peuvent donc servir à la construction où l'amélioration d'un outil de détection de fraudes.
En effet, dans un processus classique de détection des fraudes les experts en charge des réclamations effectuent des tâches chronophages pour examiner une par une les déclarations de sinistres et ensuite construire une base contenant pour chaque observation sa classe d'appartenance (frauduleuse/non-frauduleuse). Le coût opérationnel d'investigation sur une base contenant un nombre important de déclaration de sinistres peut s'avérer assez important. La réduction de ce coût pourrait donc passer par l'utilisation d'une technique de détection d'anomalies afin d'identifier les observations les plus succeptibles d'être des fraudes donc celles à investiguer.

Un des enseignements de cette étude est qu'une approche mixte pourrait s'avérer assez efficace dans le processus de détection des fraudes.
En effet, les techniques supervisées s'inspirent des comportements frauduleux passés, elles s'avèrent donc assez efficace pour modéliser un type de fraude déjà observé. Cependant, un type de fraude jamais détecté ne pourra pas être identifié par le modèle puisqu'il ne l'aura jamais appris, l'approche non-supervisée peut s'avérer efficace dans ce cas de figure car elle ne présuppose aucun étiquetage des données il existe donc une complémentarité entre les 2 approches.
Dans notre étude une approche mixte semble tirer son épingle du jeu, il s'agit de l'implémentation du modèle Random Forest sur la base rééquilibrée par l'algorithme SMOTE et enrichie par la variable explicative liée au score d'anomalies de l'Isolation Forest. Ce modèle semble particulièrement performant à notre problématique car il arrive à détecter 85% des transactions frauduleuses réelles et il obtient un score F1 de 88% à titre de comparaison l'Isolation Forest tout seul obtient un score F1 de 28% sur la même base de test.

Il est important de noter que dans la réflexion sur l'approche mixte il faudrait veiller à ne pas exposer l'algorithme aux cas suivants :

— surapprentissage : l'outil ne serait pas efficace sur d'autres types de données, ce cas de figure est observé lorsqu'on base la reflexion uniquement sur les performances,
— dénaturer l'outil : calibré l'outil pour détecter au mieux les fraudes déjà observées perdant ainsi potentiellement sa capacité à détecter d'éventuelles nouvelles structures de fraudes.

Les résultats de l'Isolation Forest sur la détection de fraudes permettent de répondre à la problématique concernant la pertinence des clusters réalisés lors de la détection d'anomalies sur la base d'arrêt de travail. Un axe d'amélioration de ce projet serait de prendre en compte dans l'étude l'impact des risques psychosociaux sur l'absenteisme.
Pour compléter cette étude, il serait intéressant d'appliquer d'autres algorithmes d'apprentissage non-supervisé tels que le LOF [2], le DBSCAN [3] ou encore le K-Prototype afin de confronter leurs performances à l'algorithme d'Isolation Forest.
Une autre possiblité d'amélioration envisageable consisterait à travailler sur ces méthodes avec des personnes ayant une expertise sur la sinistralité en question afin d'affiner les résultats.

Pour finir, la data-science offre une multitude de possibilités, l'idée est de les mettre en place et les tester pour s'assurer de leurs pertinences.
Les résultats ne sont pas toujours ceux espérés, en effet ils ne sont pas liés à la sophistication des algorithmes ou idées développées car en data-science le plus important ce sont les données utilisées, la phase de preprocessing est donc la plus importante.

2. Local Outlier Factor
3. Density-Bases Spatial Clustering of Applications with Noise

Bibliographie

Mémoires d'Actuariat

[1] Remi Bellina. Méthodes d'apprentissage appliquées à la tarification non-vie. Master's thesis, ISFA, 2014.

[2] LANGJAHR D DOMECQ G. Méthodes de provisionnement du maintien en incapacité des contrats dits franchises courtes. Master's thesis, CEA, 2013.

[3] Remi Gauville. Projection du ratio de solvabilité : des méthodes de machine learning pour contourner les contraintes opérationnelles de la méthode des sds. Master's thesis, EURIA, 2017.

[4] SANCHEZ J. Lien entre arrêt de travail et consommations médicales : application à la fraude. Master's thesis, ISFA, 2020.

[5] Karine Nguyen. Méthodes de provisionnement du maintien en incapacité des contrats dits franchises courtes. Master's thesis, IRM, 2018.

[6] LE HO Thomas. Reporting santé : comment le perfectionner à l'aide de la data science ? Master's thesis, ISFA, 2018.

Articles

[7] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof : identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[9] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[11] Jinchao Ji, Tian Bai, Chunguang Zhou, Chao Ma, and Zhe Wang. An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, 120 :590–596, 2013.

[12] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Ressources Internet

[14] CentraleSupelec. Evaluez les performances d'un modele de machine learning. https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning, 2020.

[15] Cour des Comptes. Rapport de la cour des comptes sur l'évolution des dépenses d'indemnités journalières. https://www.igas.gouv.fr/IMG/pdf/RdD2017-evolution-depenses-journalieres-web-4.pdf, 2019.

[16] Scott Fortmann-Roe. Understanding the bias-variance tradeoff. http://scott.fortmann-roe.com/docs/BiasVariance.html, 2012.

[17] Thomas C Redman. Seizing opportunity in data quality. *MIT Sloan Management Review. https ://sloanreview. mit. edu/article/seizing-opportunity-in-data-quality. November*, 29, 2017.

[18] Scikit-learn. Choosing the right estimator. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.

[19] Lilian Weng. Autoencoder. https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html.

Annexe

Méthodes d'apprentissage non-supervisé

1. Extension de l'algorithme K-MEANS : K-PROTOTYPES

1.1 Cadre théorique

Cette section n'a pas pour but de redémontrer tous les résultats de statistique, la plupart des notions utilisées proviennent de [11] .
Des solutions peuvent être mises en place afin de traiter le mélange entre variables qualitatives et variables quantitatives :

- Rendre toutes les variables qualitatives en discrétisant (découpage en classes) les variables quantitatives
- Rendre toutes les variables quantitatives à l'aide d'une AFCM

Ces méthodes ont montré leurs limites surtout au niveau de l'interprétation en sortie de modèles.
L'algorithme K-PROTOTYPES est un mixte de deux algorithmes, il combine à la fois les notions de « moyenne » et de « mode » utilisées respectivement dans les algorithmes de K-MEANS et K-MODES.
Considérons un jeu de données D contenant m variables dont p variables quantitatives et m-p variables qualitatives. Soit 2 observations X et Y appartenant à D base de données décrites par les variables

$$A_1^N, A_2^N, ... A_p^N, A_{p+1}^c, ... A_m^c.$$

La mesure de dissimilarité entre X et Y peut être mesuré par :

$$d(X,Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

où

$$\delta(x_s, q_s) = \{^{0, x_s = q_s}_{1, sinon}$$

Le premier membre de l'équation à droite est le carré de la distance Euclidienne mesure utilisée pour quantifier les écarts entre variables quantitatives, tandis que le second terme mesurant les écarts entre variables quantitatives est la distance de Hamming utile pour mesurer la dissimilarité entre deux variables catégorielles. Le terme mesure l'importance accordée au variables qualitatives. Plus γ est « grand » plus les variables qualitatives auront un poids important dans la modélisation, cependant en choisissant γ petit on réduit l'impact des variables catégorielles dans notre base.

1.2 Illustration du calcul de la distance entre deux variables mixtes

Soit A et B deux observations décrites par les variables X_1,X_2,X_3,X_4. X_3,X_4 deux variables catégorielles prenant les valeurs t,u,v,w. X_1,X_2 deux variables numériques prenant des valeurs dans l'ensemble des nombres réels. La mesure de dissimilarité pour A (1, 0, t, u), B (0, 1, t, w) et $\gamma = 1$

$$D(A,B) = (0-1)^2 + (1-0)^2 + 0 + 1 = 3$$

Etapes de la mise en place de l'algorithme

1. **Initialisation des centroïdes des K CLASS.**
   On choisit au hasard K points de la base en qualité de centroïdes initiaux.

2. **Calcul de la mesure de dissimilarité et assignation**
   On utilise l'équation (i num de l'équation plus haut) pour calculer la distance entre chaque observation X_i de notre data-set et les K-centroïdes. En fonction du résultat du calcul précédent X_i est assigné au cluster le plus proche (celui dont le centroïde a la mesure de dissimilarité la plus faible par rapport à X_i).

3. **Mise à jour du centroïde**
   En fonction des clusters ainsi formés, une mise à jour des différents centroïdes est effectuée.
   Sélection de toutes les observations dans un cluster donné pour chaque variable catégorielle le mode (valeur la plus fréquente) est choisi. En ce qui concerne les variables quantitatives la moyenne est choisie comme coordonnée du centroïde.

4. **Réassignation**
   Répétition des étapes 2 et 3 jusqu'à atteindre le minimum de la fonction coût E

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(X_i, Q_l)$$

1.3 Zoom sur la fonction de coût

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(X_i,Q_l)$$

Q_l est le vecteur représentatif des centroïdes dans le cluster l.
y_{il} un élément d'une matrice par bloc avec les propriétés suivantes :

$$0 \leq y_{il} \leq 1 \text{ et } \sum_{l=1}^{k} y_{il} = 1$$

d est la mesure de dissimilarité.

1.4 Représentation en 2 dimensions des clusters réalisés via l'algorithme K-PROTOTYPES sur la base Arrêt de travail

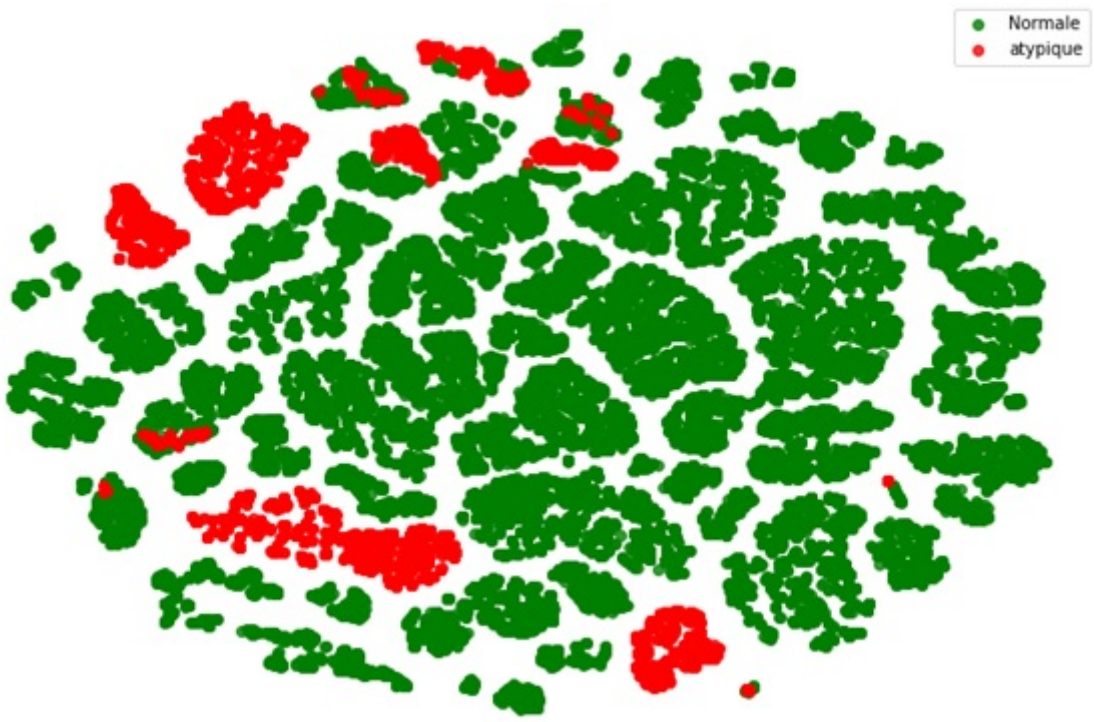


FIGURE 5.19 – Vision en 2D de la segmentation via l'algorithme K-Prototype

2. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme d'apprentissage non-supervisé introduit en 1996 [10] dont l'objectif est de diviser les points en clusters en utilisant comme règle de partitionnement la densité locale.

Quelques notions

Epsilon voisinage

L'epsilon-voisinage d'un point p $V_\epsilon(p)$, est l'ensemble des points q du jeu de données D situés dans la boule ouverte centrée en p et de rayon ϵ :

$$V_\epsilon(p) = \{q \in D \mid d(q,p) < \epsilon\}$$

Epsilon-voisinage dense

Un epsilon-voisinage est dit dense si :

$$|V_\epsilon(p)| \geq \min Pts$$

Avec ,

$\min Pts$ le nombre minimum de points pour qu'un voisinage soit désigné dense.

Directement accessible par densité

Un point q du jeu de données D est directement accessible par densité depuis un autre point p si :

- $V_\epsilon(p)$ est dense
- q appartient $V_\epsilon(p)$

Accessible par densité

Un point p est densité-accessible depuis un point q s'il existe une séquence ordonnée de points(p_1,. . . ,p_n) tel que :

- p_1=p et p_n=q
- p_{i+1} est directement accessible par densité depuis p_i pour tout i dans [2,n]

Point central, Point bordure, Point bruit

Un point est dit central si son voisinage est dense.
Un point est bordure s'il a moins de min Pts points dans son voisinage mais est dans le voisinage d'un poit central.
Un point est bruit s'il n'est ni central ni bordure.

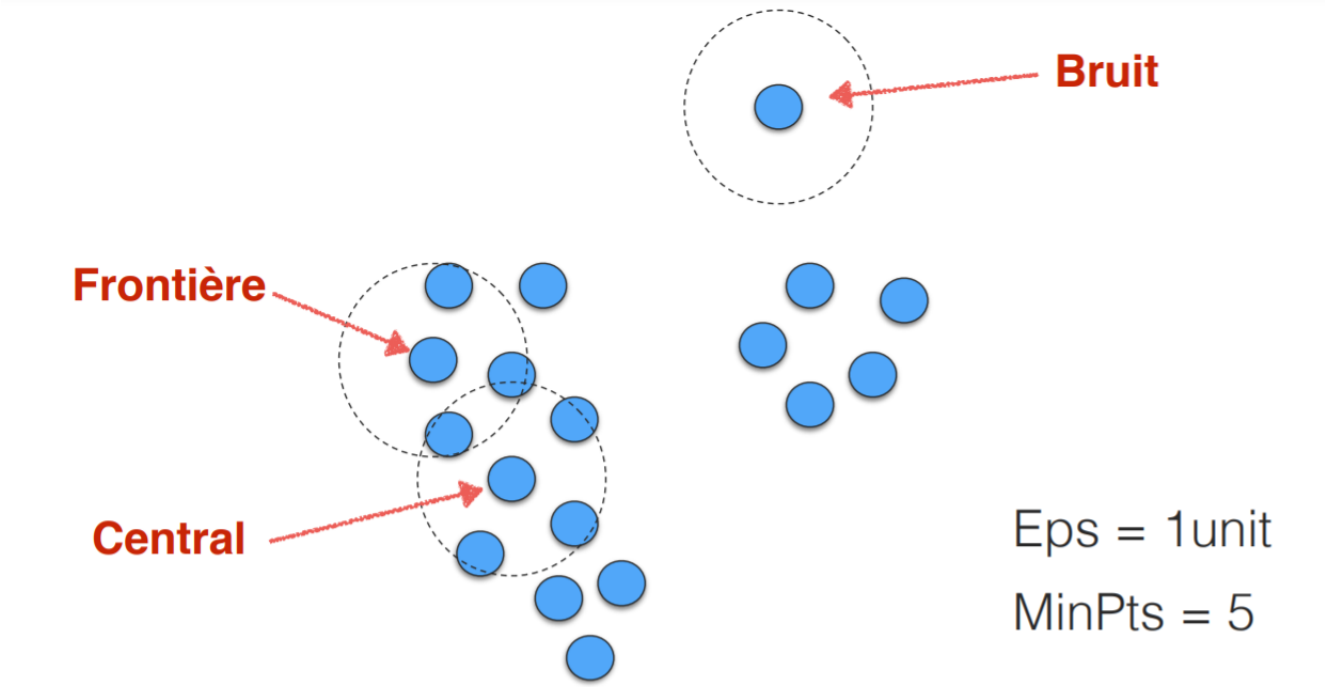


FIGURE 5.20 – Illustration de la notion de point central, bordure et bruit

La mise en place de l'algorithme respecte les principes suivants :

— Deux points centraux qui sont à moins de Eps unités l'un de l'autre sont dans le même cluster
— Tout point central qui est dans le voisinage d'un point central appartient au même cluster que lui

L'algorithme de DBSCAN est le suivant :

Etapes de la mise en place de l'algorithme

1. Prendre un point x qui n'a pas été visité

2. Construire N = voisinage (eps, x)

3. If |N| < *min*Pts:

Marquer x comme bruit
Else:

Initialiser C = {x}
Agrandir_cluster (C, N, eps, *min*Pts)

Ajouter C à la liste des clusters
Marquer tous les points de C comme visités

4. Répéter 1-3 jusqu'à ce que tous les points aient été visités

La procédure **Agrandir_cluster** est la suivante :

Agrandir_cluster (C, N, eps, *min*Pts) :

For u in N :

If u n'a pas été visité :

N'= N (eps, u)

If |N'| >= *min*Pts :

N = union (N, N')

If u n'appartient à aucun autre cluster :

Ajouter u à C

Les avantages de cet algorithme sont :

— forme des clusters libre,
— efficace en temps de calcul,
— insensible au bruit.

Les incovénients :

— sensible au fléau de la dimension,
— résultat très dépendant du choix des paramètres ϵ et $\min Pts$.

3. LOF

Les principes de la méthode Local Outlier Factor sont présentés ici [7]. Il compare la densité locale d'un point avec celles de ses k plus proches voisins cette mesure se fait au travers d'un score d'anomalies représentant le degré d'aberrance d'une observation. Les observations dont le degré d'aberrance est largement supérieur à un certain seuil sont considérées comme "anormales". La formule de calcul du score d'anomalies est la suivante :

$$\frac{\sum_{o \in N(p,k)} \frac{\mathrm{lrd}_k(o)}{\mathrm{lrd}\, k(p)}}{k}$$

Avec,
$N(p,k)$ l'ensemble des k plus proches voisins de p.
$\mathrm{lrd}_k(p)$ est la densité d'accessibilité locale de p et correspond à l'inverse de la distance d'accessibilité moyenne entre p et ses k plus proches voisins , $\mathrm{lr}\, d_k(p) = \frac{k}{\sum_{o \in N(p,k)} reachdist_k(p,o)}$.
La distance d'accessibilité de p par rapport à o considérant les k plus proches voisins ($reachdist_k(p,o)$) entre deux points (p et o), correspond au maximum entre la distance entre le k^{ieme} plus proche voisin de o ($k - \mathrm{distance}(o)$) et la distance entre p et o $d(p,o)$, soit : $reachdist_k(p,o) = \max(k - \mathrm{distance}(o)\ ,\ d(p,o))$.

Les avantages de cet algorithme sont :

- l'approche locale,
- non dépendant au calcul d'un barycentre,
- tient compte de la densité des points.

Les incovénients :

- difficulté au niveau de l'interpretation,
- choix du paramètre k des k plus proches voisins , un k faible revient à avoir des résultas précis mais instables contrairement à un k élevé qui aura pour risque de masquer des informations locales,
- choix du seuil de déclaration d'une observation comme anormale.

4. Liste des variables utilisées sur la base arrêt de travail

Variable	Typologie	Description
EX_SURVENANCE	Numérique	L'exercice de survenance
D_DARRET	Date	La date de l'arrêt
M_ARRET	Qualitative	Mois de l'arrêt
RUPCNT	Qualitative	Indiquant s'il y a eu rupture de contrat
DB_END_RUPT	Numérique	Nombre de jours depuis la rupture de contrat
FRANCHISE_RECALCULEE	Numérique	La franchise
DUREE_INDEMN_JOURS	Numérique	La durée d'indemnisation
DUREE_INDEMN_MOIS	Numérique	Durée d'indemnisation en mois
AGE_ARRET	Numérique	L'âge à l'arrêt
DUREE_AT_JOURS	Numérique	Durée de l'arrêt
DUREE_AT_MOIS	Numérique	Durée de l'arrêt (en mois)
ANCIENNETE_SURVENANCE	Numérique	L'ancienneté à la survenance
NB_AT_PERS	Numérique	Nombre d'arrêts par assuré
SEXE	Qualitative	Le sexe de l'assuré
ETADOS	Qualitative	Etat du dossier (Fermé ou Ouvert)
MOTIF	Qualitative	Motif de l'arrêt
CAG	Qualitative	Centre administratif Groupe
DEPARTEMENT	Numérique	Le département de l'assuré
TOTAL_PRESTA	Numérique	Le total des prestations
PRESTA_MENSUELLE	Numérique	Les prestations mensuelles de l'assuré
MOTIF_CLOTUR	Qualitative	Le motif de cloture du dossier
TOP_INVAL	Qualitative	Indiquant si l'assuré est entré en invalidité
CENSURE	Qualitative	Indiquant s'il y a censure
TX_CADRESUP	Numérique	Le taux de cadre supérieur
TX_INTERMEDIAIRE	Numérique	La part des professions intermédiaires
TX_EMPLOYES	Numérique	La proportion des employés
TX_OUVRIERS	Numérique	La proportion des ouvriers
TX_AUTRES	Numérique	La proportion des autres professions
TX_MARCHE	Numérique	Les travailleurs se déplaçant à la marche
TX_TC	Numérique	Les travailleurs en transport en commun
TX_TP	Numérique	Les travailleurs utilisant un transport personnel
TX_SS_TRANSPORT	Numérique	Les travailleurs à leur domicile
DENSITE_MEDECINS	Numérique	La densité des medecins pour 100000

Variable	Typologie	Description
DENSITE_GENERALISTE	Numérique	La densité des généralistes pour 100000
E_VIE_NAISSANCE	Numérique	L'espérance de vie à la naissance
TX_HANDICAP1	Numérique	Proportion d'élève handicapé en élémentaire
TX_HANDICAP2	Numérique	Part d'élève handicapé au Lycée/Collège
P_AAH	Numérique	La Part des allocataires AAH
I_VIEILLISSEMENT	Numérique	L'indice de vieillissement
TX_DIVORCE	Numérique	Le taux de divorce
TX_NORME	Numérique	Le taux de famille "Standard"
TX_MONO	Numérique	Le taux de famille monoparentale
TX_RECOMP	Numérique	Le taux de famille recomposée
NBRE_PERS_MENAGE	Numérique	Le nombre de personne moyen par ménage
NIVEAU_DE_VIE	Numérique	Le niveau de vie
TX_DE_NATALITE	Numérique	Le taux de natalité
TX_DE_PAUVRETE	Numérique	Le taux de pauvreté
SUN	Numérique	Le nombre d'heures d'ensoleillement par an

5. Décomposition des indemnités journalières

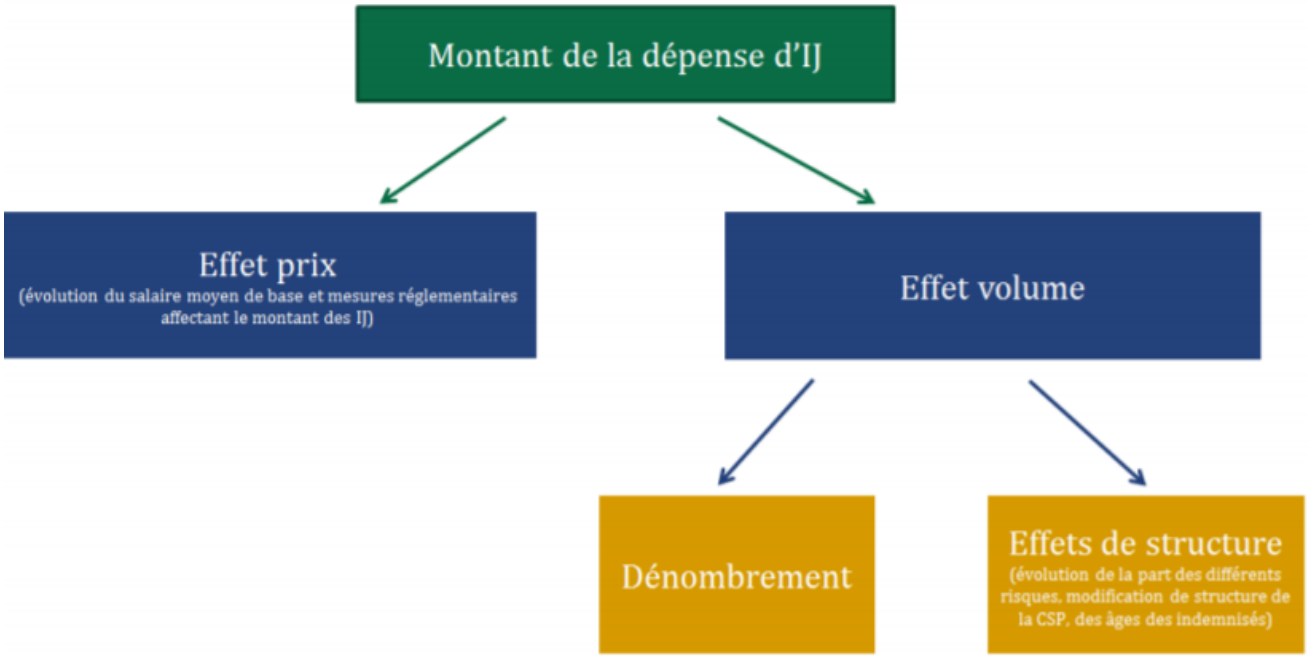


FIGURE 5.21 – Décomposition de l'évolution du montant d'indemnités journalières par la CNAM, source[15]

6. Quelques fonctions d'activations des réseaux de neurones

Fixons d'abord la variable z correspondant à la somme pondérée des entrées x_i

$$z = b + \sum_i w_i x_i$$

la fonction ReLU

La fonction Rectified Linear Unit (ReLU) est définie de la façon suivante :

$$R(z) = \max(0,z)$$

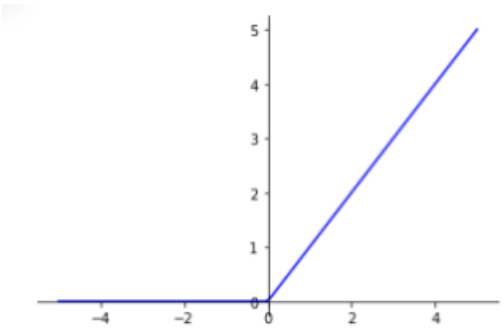


FIGURE 5.22 – Fonction ReLU

la fonction tanh

La fonction de latangeante hyperbolique (tanh) est définie de la façon suivante :

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

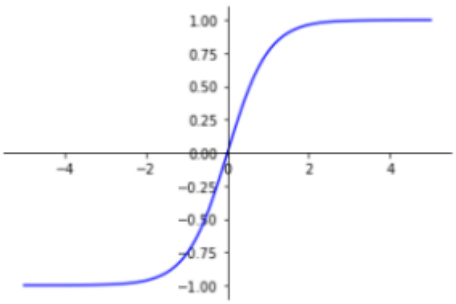


FIGURE 5.23 – Fonction tanh

la fonction Sigmoïde

La fonction sigmoïde est définie de la façon suivante :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

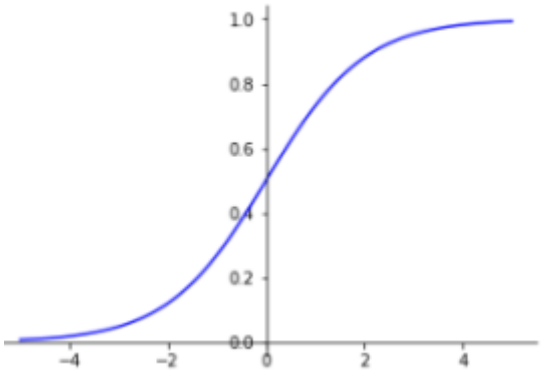


FIGURE 5.24 – Fonction Sigmoïde