Implémentation des outils de l'apprentissage automatique pour la modélisation des notations souveraines

BRIBRI Yousra **SD**

2020/2021

Résumé

L'objectif principal de cette étude serait d'effectuer des analyses et de trouver des potentiels consignes pour les preneurs de décision, concernant les différentes stratégies possibles pour améliorer la notation souveraine des pays du monde. L'obtention d'une bonne notation du crédit souverain est généralement essentielle pour les pays en développement qui souhaitent avoir accès au financement sur les marchés obligataires internationaux. Nous avons mené une démarche consistant sur l'exploitation des différentes techniques de Data Mining et de Machine Learning afin d'avoir une idée approximative sur les indicateurs susceptibles d'impacter cette notation.

Par ailleurs, nous sommes amenés à traiter la problématique suivante "**établir un modèle de prédiction de la notation souveraine du Maroc et distinguer les facteurs qui permettent d'améliorer ou qui pénalisent cette notation**". Nous avons réalisé une étude descriptive en se basant sur les données de l'agence de notation Standard & Poor's. Afin d'assurer une prévision significative de la variable Rating, nous avons consulté plusieurs études théoriques discutant ce phénomène, nous avons alors repéré les variables nécessaires afin de mener notre prévision. Une fois les variables transformées et optimisées, nous avons effectué le nettoyage de données qui est une étape cruciale pour préserver une performance algorithmique maximale. En premier lieu, nous avons modéliser la notation souveraine en prenant en considération deux catégories : **Spéculative** et **Investissement**. Ensuite, nous nous sommes intéressés aux notations détaillées : AAA, AA, A, BBB, BB, B et CCC.

Pour établir ces modélisations nous avons adopté comme modèle : **Régression Logistique** et **Support Vector Machine**.

L'entrainement de ces modèles nous a alors donné des résultats satisfaisants en termes de précision et de significativité. Ces modèles construits sont capable de différencier les notations souveraines des pays en fonction des indicateurs.

Mots Clés

Notation souveraine, Machine Learning, Binaire, Multiclass

Dédicace

Je dédie cet humble et modeste travail,
À mes chers parents **Fouzia** et **Abdelhafid** pour leur amour, leur support et leurs sacrifices, vous avez fait de moi la personne que je suis aujourd'hui et je vous en serai toujours reconnaissante.
À ma soeur **Chaimaa** qui a toujours été là pour moi, ta présence est pour moi une lueur d'espoir.
À tous mes amis que je remercie pour leur joie de vivre, pour cette belle amitié et pour leur aide.

Remerciements

A travers ce passage nous souhaitons remercier chaleureusement toute personne ayant contribuer à la réussite de notre stage.

Également nous destinons nos chaleureux remerciements à Monsieur HASSNAOUI Brahim notre encadrant au sein de la Direction des Études et des Prévisions Financière (DEPF), pour tous les efforts qu'il n'a cessé de fournir pour nous encadrer avec ses remarques enrichissantes et constructives, qui nous ont été d'un appui considérable tout au long de ce stage.

Nous remercions aussi tout le corps professionnel de l'Institut National de Statistique et d'Économie Appliquée et tout le personnel de la Direction des Études et des Prévisions Financières.Qu'ils trouvent toute l'expression de notre profonde reconnaissance car sans eux ce travail n'aurait jamais vu le jour.

Table des matières

Liste d'abréviations

IA	Intelligence Artificielle
ML	Machine Learning
DL	Deep Learning
SVM	Support Vector Machine
RL	Régression Logistique
RFE	Elimination récursive de caractéristiques
ANC	Agence de notation de crédits
SP	Standard & Poor's

Liste des figures

Liste des tableaux

Introduction

En 1909, la notation financière a vu le jour avec John Moody, fondateur de l'agence Moody's, qui remarque que l'accroissement continu du nombre d'obligations du secteur corporate émises sur les marchés financiers américains ne permet plus de distinguer le degré de risque des divers titres échangés. Il a donc l'idée de publier des fiches de renseignements économiques et financiers sur les entreprises et d'ajouter une note qui mesure le risque de défaut de chaque obligation émise par la société en question. Ces fiches ont connu un succès remarquable et Moody's voit rapidement surgir d'autres concurrents. Ceci dit, on pourra définir brièvement la notation financière comme étant l'appréciation, par une agence de notation financière, du risque de solvabilité financière d'une entreprise, d'une opération ou d'un état. La notation financière d'un état est connue sous le nom de notation souveraine : qui représente le thème de notre sujet de recherche.
Afin de mieux éclairer ces différents points, cet ouvrage commencera par présenter l'histoire de la notation souveraine née il y a plus d'un siècle aux Etats-Unis, et définira ce qu'est exactement une notation. Seront ensuite abordées la méthodologie et les différentes échelles de notation des agences spécialisées dans cette pratique.
Dans le but de répondre à notre problématique qui vise à déterminer les variables susceptibles d'impacter la notation, une revue de littérature des différentes études empiriques qui ont été faite s'est avéré nécessaire afin d'avoir une idée sur la façon avec laquelle on va traiter notre problématique ainsi que les différentes variables qu'on pourra insérer comme étant variables explicatives.
Par ailleurs, nous sommes amenés à traiter cette problématique avec beaucoup de considération pour étudier les différents indicateurs qui contribuent à cette notation. Ceci est rendu possible bien plus que jamais grâce aux différents algorithmes de Machine Learning qui maximisent le plus réellement possible le succès et la significativité des prévisions et entraînent alors de meilleures décisions. Nous allons alors exploiter nos différentes connaissances acquis lors de notre parcours académique, et mener plusieurs recherches approfondies afin de pouvoir construire un modèle qui pourra nous révéler les indicateurs qui impactent et expliquent le mieux possible notre variable cible.

Première partie

Cadre général

Chapitre 1

Notation souveraine

1 Historique de la notation souveraine

1.1 Contexte

Le risque souverain est inhérent à l'existence même des États. Déjà, au cours de l'Antiquité, des cités-États se sont montrées incapables de rembourser leurs dettes. Au cours des siècles suivants, avec la formation des États-nations, le problème de l'insolvabilité de certains États s'est encore posé. L'Espagne, pourtant première puissance mondiale, a fait défaut sur sa dette au cours du Siècle d'Or. La France et les autres grandes nations européennes ont fait de même aux XVIe, XVIIe et XVIIIe siècles. Le risque de défaut des États est donc très ancien et il a souvent entraîné la chute des banquiers prêteurs.

Jusqu'au début du XXe siècle, l'analyse du risque souverain était l'apanage des banques. Par exemple, dans les années 1810, la famille Rothschild, qui possède une bonne réputation dans le domaine bancaire et financier, accepta de prêter au roi de Prusse mais lui imposa des taux d'intérêt plus élevés qu'attendu, au motif que son royaume ne connaissait pas d'état de droit. À la fin du XIXe siècle, le Crédit lyonnais développa une méthode de notation souveraine à usage interne qui reposait en bonne partie sur le service de la dette de l'État. Il faut finalement attendre 1918 pour voir apparaître les premières analyses de risque souverain complètement indépendantes des banques : elles sont publiées par l'agence Moody's Investors Service (plus communément appelée Moody's), qui se lance dans la notation des titres obligataires souverains. Ces notes, ou ratings, reflètent une probabilité de défaut et renvoient donc à un certain niveau de solvabilité de l'État.

1.2 Historique

Le contexte dans lequel est née la notation financière mérite d'être brièvement rappelé. Les tout premiers ratings apparaissent en 1909. John Moody, fondateur de l'agence Moody's, comprend que l'accroissement continu du nombre d'obligations du secteur corporate émises sur les marchés financiers américains – majoritairement par des entreprises industrielles et des compagnies de chemins de fer – ne permet plus de distinguer le degré de risque des divers titres échangés. Il a donc l'idée de publier des fiches de renseignements économiques et financiers sur les entreprises et d'ajouter une note qui mesure le risque de défaut de chaque obligation émise par la société en question. Ces fiches sont compilées dans de volumineux manuels qui sont vendus aux fonds d'investissement, établissements de crédit et investisseurs particuliers américains. Le succès est immédiat et Moody's voit rapidement surgir trois concurrents sérieux : Standard Statistics, Poor's (qui fusionneront en 1941 pour former Standard Poor's) et Fitch.

À partir de 1915, la France et la Grande-Bretagne sollicitent les investisseurs américains afin de financer l'effort de guerre. John Moody, conscient que les États-Unis sont en train de devenir le premier pays créancier au monde, se lance alors dans la notation souveraine au début de l'année 1918. Au cours des années 1920, la majorité des États européens et latino-américains délaissent les marchés londonien et parisien et émettent des obligations d'État sur la place de New York. Tous ces États sont automatiquement notés par Fitch, Moody's, Poor's et Standard Statistics. Le krach de 1929 puis la Grande Dépression qui suit déclenchent une vague massive de défauts souverains à partir de 1931. L'incapacité de la plupart des pays d'Amérique du Sud et d'Europe centrale, et aussi de l'Allemagne, à rembourser leur dette, ainsi que les nombreuses faillites de fonds d'investissement et de banques, assèchent complètement le marché obligataire souverain. Le déclenchement de la Seconde Guerre mondiale porte le coup de grâce au développement des marchés financiers. Les modalités de reconstruction des économies européennes après 1945, qui impliquent des mesures de soutien exceptionnelles (tel le plan Marshall) puis le retour à la croissance au cours des décennies suivantes, fondé sur l'interventionnisme étatique et l'intermédiation bancaire, rendent superflu tout recours aux marchés financiers.

À partir de 1915, la France et la Grande-Bretagne sollicitent les investisseurs américains afin de financer l'effort de guerre. John Moody, conscient que les États-Unis sont en train de devenir le premier pays créancier au monde, se lance alors dans la notation souveraine au début de l'année 1918. Au cours des années 1920, la majorité des États européens et latino-américains délaissent les marchés londonien et parisien et émettent des obligations

d'État sur la place de New York. Tous ces États sont automatiquement notés par Fitch, Moody's, Poor's et Standard Statistics. Le krach de 1929 puis la Grande Dépression qui suit déclenchent une vague massive de défauts souverains à partir de 1931. L'incapacité de la plupart des pays d'Amérique du Sud et d'Europe centrale, et aussi de l'Allemagne, à rembourser leur dette, ainsi que les nombreuses faillites de fonds d'investissement et de banques, assèchent complètement le marché obligataire souverain. Le déclenchement de la Seconde Guerre mondiale porte le coup de grâce au développement des marchés financiers. Les modalités de reconstruction des économies européennes après 1945, qui impliquent des mesures de soutien exceptionnelles (tel le plan Marshall) puis le retour à la croissance au cours des décennies suivantes, fondé sur l'interventionnisme étatique et l'intermédiation bancaire, rendent superflu tout recours aux marchés financiers.

2 Agences de notation

Les agences de notation sont des entreprises privées dont l'activité principale consiste à évaluer la capacité des émetteurs de dette à faire face à leurs engagements financiers. Il s'agit bien d'organismes privés à but lucratif et non d'organismes réglementaires ou gouvernementaux. Les agences de notation sont des acteurs incontournables des marchés. En effet les notations sont souvent utilisées dans le cadre réglementaire d'une part, et aussi dans les stratégies de nombreux investisseurs.

Depuis l'apparition de la notation financière, plus de 150 agences de notations ont existé dans le monde mais trois agences dominent ce secteur. Il s'agit de Standard Poor's, Moody's et Fitch Rating, dont les parts de marché s'élèvent pour les deux premières à 40% et pour la troisième à 10%, soit 90% du marché mondial de la notation souveraine. Leur mission principale consiste à délivrer des informations sur le risque de défaut de paiement des entreprises ou des gouvernements, en notant la qualité des titres qu'ils émettent.

2.1 Moody's

Moody's est une société active dans l'analyse financière d'entreprises commerciales ou d'organes gouvernementaux. Elle est également connue pour ses notations financières standardisées des grandes entreprises en fonction du risque et de la valeur de l'investissement.

Moody's a été fondée en 1909 par John Moody, journaliste financier reconverti, qui crée la notation. La société jauge les risques des entreprises en s'appuyant sur une grille de notes, qui permet de résumer les risques pris par le créancier. La société a connu plusieurs évènements marquants et qui ont impacté l'économie mondiale. Par exemple, en 1931, l'agence qui note aussi les dettes publiques, dégrade la note de la Grèce. La république grecque qui mène alors d'importantes réformes économiques s'en trouve déstabilisée et fait face à une crise économique pendant plusieurs années. Entre-temps, en 1936, les dirigeants de Moody's expriment leur regret sur ce qui se passe et annoncent qu'ils arrêtent de noter les dettes publiques.

En 1962, Moody's Investors Service est racheté par Dun Bradstreet. En décembre 1999, Dun Bradstreet a annoncé qu'elle allait délocaliser Moody's Investors Service en une société distincte cotée en bourse.

2.2 Standard & Poor's

Standard Poor's (SP) est une filiale de McGraw-Hill qui publie des analyses financières sur des actions et des obligations. C'est une des principales sociétés de notation financière, elle est connue sur le marché américain pour son indice boursier SP 500.

L'histoire de la société remonte à 1860, avec la publication du « Manual of Railroads of United states » par Henry Varnum Poor. Ce livre a compilé des informations complètes sur la situation financière et opérationnelle des États-Unis. En 1906, Luther Lee Blake a fondé le Standard Statistics Bureau, dans le but de fournir des informations financières sur les entreprises non ferroviaires. Au lieu d'un livre publié chaque année, Standard Statistics utilisait des cartes permettant des mises à jour plus fréquentes. En 1941, Paul Talbot Babson a acheté Poor's Publishing et l'a fusionné avec Standard Statistics pour devenir Standard Poor's Corp. En 1966, la société a été acquise par The McGraw-Hill Companies, ce qui permet à McGraw-Hill de s'étender dans le domaine des services d'information financière.

2.3 Fitch Ratings

Fitch Ratings Ltd. est une agence de notation financière internationale, elle a été fondée par John Knowles Fitch le 24 décembre 1913 à New York sous le nom Fitch Publishing Company. Elle a été fusionnée avec la société IBCA Limited, basée à Londres, en décembre

1997, passant ainsi sous le contrôle du holding français Fimalac dont le principal actionnaire est le français Marc Ladreit de Lacharrière. En 2000, elle a acquis les sociétés Duff Phelps Credit Rating Co. (basée à Chicago) et Thomson BankWatch, puis s'est retrouvée exposée aux critiques lors de la crise financière de 2007 à 2011. Depuis le 12 décembre 2014, elle est détenue à 80% par le groupe Hearst.

Ces trois agences se distinguent d'abord par le statut Nationally Recognized Statistical Rating Organizations (organisations de notation statistique nationalement reconnues : NRSRO) qui leur est accordé par la SEC (Securities and Exchange Commission). Ce statut permet aux agences d'agir sur le marché américain et d'avoir une visibilité mondiale. D'autre part, la majorité des investisseurs s'appuient sur les notations de ces agences à cause de leur certification et obtention du statut NRSRO.

D'autres agences de notation ont été créées en Chine, Russie, Japon...sans pour autant obtenir le statut NRSRO tel que Dagong Global Credit Rating, agence de notation financière chinoise, fondée en 1994. Bien que ses notations aient été jugées plus crédibles que celles de ses concurrentes, l'agence chinoise n'a pas réussi à obtenir l'accréditation de la SEC en octobre 2010.

3 Présentation de la notation souveraine

La notation financière souveraine consiste à mesurer la probabilité de défaut des états souverains. Elle cherche à évaluer la capacité d'un état à payer ses obligations relatives à sa dette commerciale vis-à-vis des créditeurs non officiels. En fournissant une évaluation du risque souverain, la notation se focalise exclusivement sur la solvabilité du gouvernement central. Les autorités d'un pays recherchent généralement des notations des agences accréditées afin de leur faciliter l'accès de leur gouvernement ainsi que d'autres émetteurs à l'intérieur du pays, aux marchés internationaux des capitaux où de nombreux investisseurs préfèrent les titres notés aux titres non notés. Ceci dit, une note est une indication de la probabilité de défaut, de telle sorte qu'il y ait une correspondance entre la note et la probabilité de défaut.

Les notes attribuées par les agences de notation sont des opinions indépendantes fondées sur des arguments ainsi qu'une méthodologie propre à chaque agence. L'approche de l'analyse du risque de crédit se fait via des jugements quantitatifs et qualitatifs qui captent la volonté et la capacité des états à honorer leurs obligations.

3.1 Méthodologie de la notation souveraine

La notation financière souveraine est une appréciation subjective du risque souverain basée sur une combinaison d'indicateurs qualitatifs et quantitatifs dans un modèle de notation. Dans cette notation financière, on fait toujours la distinction entre la notation de l'émetteur et la notation de la dette. La première notation indique l'état général de la solvabilité financière d'un émetteur souverain. La deuxième indique la solvabilité spécifique d'un émetteur souverain par rapport à un instrument financier spécifique. L'objet de l'étude qu'on va mener est la notation de l'émetteur ou plus généralement la notation de tout un pays.

A l'analyse des notes méthodologiques des différentes agences de notation souveraine, on constate que les données utilisées dans les modèles relèvent de cinq piliers : environnement politique et institutionnel, performance macro-économique, état des finances publiques, système financier et monétaire et finances extérieures :

- Au niveau de l'analyse de la performance macro-économique, l'agence cherche à appréhender les caractères saillants des structures économiques du pays et à évaluer la cohérence des politiques suivies, compte tenu des contraintes auxquelles il est soumis. Le but est d'examiner si le pays dispose d'une croissance économique capable de générer une base de revenus et des marges de manœuvre budgétaires et monétaires plus importantes pour qu'il puisse honorer ses engagements financiers.
- À travers l'analyse politique et institutionnelle, l'agence cherche à mesurer les risques liés à l'environnement politique et institutionnel, susceptibles d'impacter la volonté d'un État à payer sa dette extérieure.
- L'analyse des finances publiques se fait à travers l'analyse des comptes de l'Etat afin d'évaluer le niveau des ressources générées par l'activité économique et s'assurer qu'il pourra assurer le service de sa dette commerciale.
- Au niveau des comptes extérieurs et vulnérabilités externes, l'agence cherche à mesurer le degré d'intégration et de dépendance du pays vis-à-vis de l'extérieur en matière des échanges des biens et des capitaux et à évaluer les vulnérabilités et risques externes qui en résultent. Le point saillant de cette analyse et d'évaluer la capacité de l'économie à générer suffisamment de devises pour pouvoir payer le service de la dette extérieure. L'exercice consiste en l'évaluation du passif en devises de l'ensemble de l'économie au

regard des réserves de change disponibles et des flux futurs.

- Sur le plan du système monétaire et financier, l'agence analyse la solidité du système bancaire et la crédibilité de la politique monétaire et le statut de la banque centrale.

Enfin, il est important de noter que l'analyse des performances réalisées dans le passé doit toujours être complétée par une analyse prospective par la réalisation de prévisions de moyen terme concernant les agrégats macro-économiques et d'autres indicateurs économiques, car après tous ce qui intéresse le plus c'est d'appréhender la probabilité de défaut dans le futur.

3.2 Les échelles de notation et leurs significations

Afin de déterminer la solvabilité des différents émetteurs, les agences utilisent une échelle de notation spécifique, qui permet facilement une distinction entre les émetteurs des obligations les plus solvables.
Une échelle de notation se traduit par des lettres de A assortis de + ou de − à D, chaque lettre détermine le degré de risque de défaut et la qualité de l'émetteur, ces symboles allient des ratings les moins risqués (AAA), donc l'assurance maximale d'être remboursée, aux plus risqués (C ou D) correspondant à un défaut partiel ou total. Lorsque l'on décline l'alphabet, la qualité de crédit se dégrade. Chaque échelle est subdivisée en deux catégories : La **catégorie d'investissement** (investment grades) regroupe les notes de meilleures qualités de crédit présentant moins de défaut de paiement et la **catégorie spéculative** (speculative grades) regroupe des notes de qualité de crédit mauvaise ou médiocre exposées au risque de défaillance.

Chaque catégorie est subdivisée en notation à long terme et notation à court terme, une forte corrélation existe entre les notes à court et à long terme. Les notes à long terme vont de (AAA) pour la meilleure, à (BBB-) dans la catégorie investissement, et de (BB+) à (D) pour la plus mauvaise catégorie spéculative.

Une note permet donc un classement en fonction des caractéristiques particulières du titre miné et des garanties offertes par son émetteur ; chaque symbole correspond à une graduation sur l'échelle de notes.

D'où, l'échelle de notation, sert comme outil de classement et fait comprendre qu'un caractère alphanumérique ou un symbole n'est pas le choix du hasard. Chaque note est la

traduction d'une situation spécifique liée au risque de défaut d'un emprunteur, de ce fait, l'organisme qui délivre les notes doit toujours fournir une grille de lecture de la note de crédit.

Le tableau ci-dessous explique la signification des différentes notations utilisées par les deux agences de notation Moody's et Standard Poor's, à noter que l'agence Fitch, utilise la même grille que SP.

S&P	Moody's	Définition des symboles de notation des émetteurs
AAA	Aaa	**S&P** Capacité extrêmement forte à respecter ses engagements financiers. **Moody's** Sécurité financière exceptionnelle. Même en cas de changements de la situation financière, sa position restera fondamentalement forte.
AA	Aa	**S&P** Capacité très forte à respecter ses engagements financiers. Il diffère faiblement de la précédente notation. **Moody's** Excellente Sécurité financière. Il est moins bien noté que Aaa car le risque à long terme apparait supérieur. Ces deux notes constituent des émetteurs à haut grade.
A	A	**S&P** Capacité forte à respecter ses engagements financiers. Plus susceptible d'être affecté par les changements de circonstances et des conditions économiques que les précédentes notes. **Moody's** Bonne sécurité financière. Des éléments actuels peuvent suggérer une possibilité de dégradation dans le futur.
BBB	Baa	**S&P** Capacité adéquate à respecter ses engagements financiers. Des changements défavorables de circonstances ou de conditions économiques vont vraisemblablement affaiblir sa capacité à respecter ses engagements financiers. **Moody's** Sécurité financière adéquate. Mais certains éléments protecteurs peuvent manquer ou être incertains sur une longue période.
BB à C	Ba à C	Ces notes sont considérées « spéculatives ». Alors que les précédentes sont considérées « investissement ».
BB	Ba	**S&P** De grandes incertitudes et risques face aux mauvaises conditions économiques et financières peuvent mener à une capacité inadéquate de respecter ses engagements financiers. Moins vulnérable sur le court terme que les notations plus basses. **Moody's** Sécurité financière incertaine. Souvent la capacité de cette émetteur à respecter ses engagements financiers est modérée et incertaine dans le futur.
B	B	**S&P** Capacité de respecter ses engagements financiers sur le court terme. Des conditions d'activité, financière ou économique, défavorables vont vraisemblablement détériorer sa capacité ou sa volonté de respecter ses engagements financiers. **Moody's** Sécurité financière pauvre. L'assurance du respect de ses engagements financiers sur une longue période est faible.
CCC	Caa	**S&P** Actuellement vulnérable. Sa capacité de respecter ses engagements financiers dépend de conditions d'activité, financière et économique, favorables. **Moody's** Sécurité financière très pauvre. Ils peuvent être en défaut ou des éléments de risques présents peuvent empêcher le respect des remboursements prévus.
CC	Ca	**S&P** Actuellement hautement vulnérable. **Moody's** Sécurité financière extrêmement pauvre. Souvent en défaut ou des faiblesses importantes.
C	C	**S&P** Hautement vulnérable à la cessation de paiement. **Moody's** Habituellement en défaut et le potentiel de recouvrement faible.
D	D	En défaut sur une ou plusieurs de ses obligations financières.

TABLE 1.1 – Symboles et définitions de notes utilisés par Standard & Poor's et Moody's

4 Revue de littérature sur la notation souveraine

4.1 Quelques études empiriques

Il existe une vaste littérature traitant les déterminants de la notation financière souveraine. Les études menées dans ce cadre peuvent être classées en trois catégories : la première correspond aux études qui se sont intéressées à l'examen de l'impact des variables quantitatives. La deuxième concerne celles qui se sont préoccupées de l'impact des variables qualitatives et la dernière est un mixte entre les deux premiers courants et/ou traite cette relation soit dans le cadre d'un contexte précis (groupe de pays, pays en particulier, continents, etc.), soit sous un angle donné (grade de notation et asymétrie dans la notation, impact de la crise, relation de court ou de long terme. . .).

Au niveau de la première catégorie, on trouve l'étude précitée de Cantor Packer (1996) qui avaient retenu huit variables potentiellement explicatives de la notation financière souveraine accordée par Moody's et SP. Leur étude suggère que, dans une large mesure, les notations souveraines peuvent s'expliquer par un petit nombre de critères bien définis : revenu par habitant, croissance du PIB, inflation, dette extérieure, niveau de développement économique et historique des défauts de paiement. Les auteurs n'ont pas trouvé de relation systématique entre les notations et le déficit budgétaire ou le déficit courant de la balance des paiements courants. Afonso, (2002) a examiné l'impact de six variables parmi les huis variables retenues par Cantor Packer. L'auteur a abouti aux mêmes conclusions. Les six variables considérées ont été significativement corrélées aux notations financières accordées par Moody's et SP : revenu par tête, dette extérieure, niveau de développement, historique de défauts de paiement, taux de croissance économique et taux d'inflation. En se basant sur le travail de Cantor Packer, Canuto, Santos, Porto (2012) ont repris les mêmes variables que les études précédentes, mais sous une forme un peu différente. Par exemple au lieu de considérer le compte courant/PIB, ils ont utilisé la variable (imports + exports de biens et services)/PIB ; pour la dette, ils ont retenu la dette totale brute de l'administration centrale/les recettes fiscales. L'étude a confirmé la significativité des variables choisies, tout en soulignant la prépondérance de la dette publique/les recettes ordinaires, la dette extérieure/les recettes courantes des échanges extérieurs des biens et services, l'ouverture commerciale et le taux de croissance économique.

Au niveau de la catégorie des études relatives à l'impact des variables qualitatives sur la note financière souveraine, on trouve celle de Butler Fauver, (2006). Les auteurs ont

constaté que la qualité des institutions juridiques et politiques d'un pays, tel que mesurée par la voix des citoyens, le contrôle de la corruption, la stabilité politique et le rôle de la loi, joue un rôle essentiel dans la détermination de cette note. Une augmentation d'un écart-type de l'indice de l'environnement juridique et politique se traduit par une augmentation moyenne de la note de crédit de 0,466 écarts-types, même lorsque les facteurs quantitatifs, tels que le PIB par habitant, l'inflation, la dette extérieure par PIB, les défauts antérieurs et le niveau de développement sont contrôlés. Dans une étude similaire, Ozturk (2014) a montré également que la qualité institutionnelle avait un impact positif sur la notation financière souveraine. L'auteur a retenu les mêmes variables que Butler Fauver, (2006) en y ajoutant l'effectivité du gouvernement et la qualité des textes juridiques. Selon ces auteurs, ces deux dernières variables expliquent significativement les faibles notations souveraines.

S'agissant de la troisième catégorie des études on peut citer Afonso, Gomes, Rother (2010) qui ont retenu dans leur étude des variables à la fois quantitatives et qualitatives. Ces auteurs ont trouvé que ces variables n'exercent pas, de la même manière, leurs effets sur la notation financière. Ils font une distinction entre les variables qui exercent un effet de court terme sur le rating et les variables qui ont un effet de long terme. Dans le premier groupe de variables on trouve la variation du revenu par tête, le taux de croissance économique et le solde budgétaire. Dans le deuxième, on trouve l'effectivité du gouvernement, la dette extérieure, les réserves de change et l'historique de défaut. Dans une étude relative aux déterminants de la note souveraine de SP accordées aux pays émergents, Erdem Varli (2014) ont constaté que les facteurs les plus significatifs sont le solde budgétaire/PIB, le PIB par habitant, les indicateurs de gouvernance et les réserves de change/PIB.

En reprenant les variables de Cantor et y ajoutant trois autres, à savoir, le taux de change réel, le taux de chômage et le coût de l'unité de travail, Bissoondoyal-Bheenick (2005) fait constater que l'analyse des indicateurs économiques et financiers ne peut aider à comprendre les déterminants de la notation financière et ce, parce que « les mesures quantitatives fournissent des informations sur les performances passées de l'économie et sur ses caractéristiques structurelles fondamentales. Ils sont essentiellement rétrospectifs, tandis que l'analyse de la notation souveraine nécessite des évaluations prospectives du risque de défaut sur un horizon de temps moyen à long terme. Par conséquent, l'examen des performances passées doit être complété par des projections à moyen terme et par l'élaboration d'une série de scénarios qui testent la vulnérabilité de la situation économique, politique et financière d'un pays à une variété de chocs générés tant à l'intérieur qu'à l'extérieur ». Aussi, les variables économiques n'ont pas la même importance pour les pays qui sont bien notés avec une longue

histoire de stabilité financière que pour les pays mal notés qui connaissent encore des changements structurels comme c'est le cas de la majeure partie des pays émergents.

Après avoir rajouté d'autres variables quantitatives et qualitatives relatives, à savoir le ratio de l'investissement sur le PIB, l'indice de la corruption, la qualité de la réglementation, la reddition des comptes, et le rôle de la loi et stabilité politique à leur modèle, Mellios Paget-Blanc (2006) ont trouvé que les notations souveraines sont principalement influencées par le revenu par tête, le solde budgétaire, les variations du taux de change réel, le taux d'inflation, l'historique des défauts et la corruption.

Pour Jaramillo (2010), les pays à économie émergente recherchent activement le statut de l'«Investment Grade » afin d'élargir le nombre d'investisseurs étrangers potentiellement intéressés par leurs titres de dettes et de financer leur budget et économie à moindre coût sur le MFI. Dans son étude sur les déterminants de l'attribution de ce « grade », l'auteur constate qu'une poignée de variables explique significativement son attribution : la dette publique extérieure, la dette intérieure, le risque politique, les exportations, et la profondeur du système financier.

Dans une étude similaire, Broto Molina (2016) ont tenté d'approcher l'impact différencié des différentes variables sur la notation souveraine de SP en faisant la distinction entre l'«Investment Grade » et le « Speculative Grade ». Leurs résultats indiquent que les fondamentaux économiques pourraient aider à aplanir la trajectoire de la notation en évitant le déclassement au « Speculative Grade », mais ils n'ont pas l'effet similaire une fois l'État est classé « Speculative Grade ». En d'autres termes, le fait d'avoir des fondamentaux domestiques sains peut inciter l'agence de notation à modifier la trajectoire de la dégradation, de sorte que les autorités d'un pays disposent d'un instrument pour lisser les dégradations. Cependant, les « upgardes » sont d'une nature assez différente, dans la mesure où les pays dégradés ont peu de possibilités à regagner dans le futur l'«Investment Grade », même en enregistrant une reprise spectaculaire. Une fois dégradé, un État est dans une sorte de trappe à laquelle il n'est pas facile d'échapper.

4.2 Analyse des déterminants de la note financière souveraine du Maroc

Après avoir énoncé les différentes études empiriques concernant la notation souveraine ainsi que les variables explicatives. On passe à une autre étude consacrée spécialement à la notation financière du Maroc, faite par deux enseignants chercheurs au laboratoire de l'économie et management des organisations à l'université Ibn Tofail à Kénitra, et comme sujet de l'étude l'analyse des déterminants de la note financière souveraine du Maroc.

En effet, la corrélation existant entre les notes de crédits souverains accordées par les agences de notation de crédits (ANC) et les conditions d'accès aux capitaux étrangers fait que les autorités marocaines prêtent une attention particulière aux notes obtenue par le Maroc. La compréhension des déterminants de la note souveraine du Maroc est critique au vu de ses implications en termes de politiques économiques et de réformes structurelles. Dans cette perspective, cette étude examine dans quelle mesure la note souveraine du Maroc reflète la performance enregistrée par ses fondamentaux économiques et rend compte de son risque souverain. L'objectif est d'identifier un nombre limité de facteurs qui peuvent faire l'objet de l'essentiel de l'action publique afin de réduire le risque souverain. L'étude a retenu le cas de la note financière souveraine accordée par l'agence Standard Poor's (SP).

La première étape de l'étude était l'identification des variables à retenir au niveau de l'analyse empirique. Vingt variables, couvrant les cinq piliers de notation financière considérés par l'agence SP, ont été considérées, à savoir : la performance économique, le développement institutionnel et la stabilité politique, les finances publiques, les comptes extérieurs, et le secteur monétaire et le développement financier. Les prix du phosphate sur les marchés internationaux a été retenue parmi les variables explicatives.

Ensuite, le modèle ARDL a été introduit pour mener l'analyse de la relation entre la variable dépendante, qui est la notation souveraine accordée par l'agence SP au Maroc, et ses variables explicatives potentielles. Une transformation numérique des notes alphabétiques a été effectuée afin de pouvoir procéder à une estimation paramétrique à l'aide d'un modèle ARDL. Les données afférentes à ces variables sont mensuelles et couvrent la période 1998-2019. Le test de stationnarité a révélé que toutes les variables sont stationnaires, et les différents tests de diagnostic ont montré que notre modèle est robuste et offre une estimation BLUE à nos coefficients.

Un nombre limité de variables (neuf) se sont avérées significatives à expliquer la note souveraine accordée par SP au Maroc : la croissance économique, l'inflation, le contrôle de la corruption, la reddition des comptes, la profondeur du système financier, le déficit budgétaire, la dette total du Trésor, la dette extérieure du Trésor et les avoirs extérieurs nets. Les résultats confortent ainsi les conclusions des études empiriques précédentes, tout en faisant émerger pour la première fois le pouvoir explicatif significatif de la note financière souveraine du Maroc de trois variables : les avoirs extérieurs nets, la dette totale du Trésor et le niveau de développement financier du Maroc.

La relation de la notation financière accordée au Maroc par SP avec ses variables explicatives est majoritairement une relation de court terme. Seules les variables des finances publiques paraissent être liées à la note souveraine par une relation de long terme. Il s'agit du solde budgétaire et de la dette totale du Trésor.

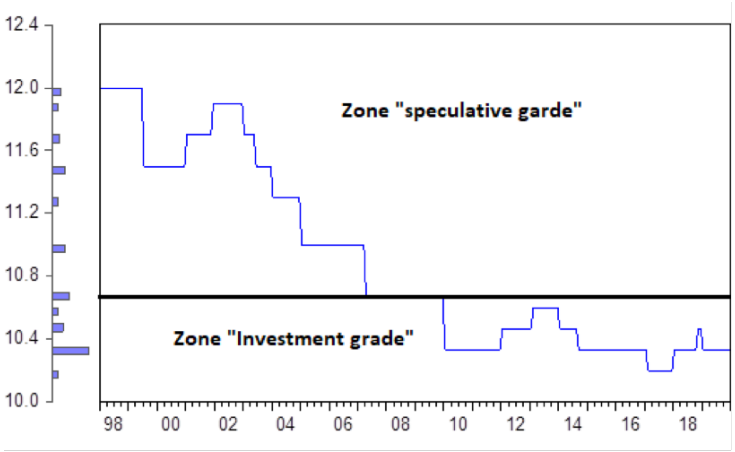

FIGURE 1.1 – Note moyenne souveraine du Maroc (1998-2019)

En outre, l'étude montre que le Maroc a dû réaliser des performances exceptionnelles, surtout au niveau des finances publiques, pour quitter la zone du « Speculative Grade » en mars 2010 et acquérir le statut de l'« Investment Grade ». Les implications de ces résultats en termes de politiques économiques sont importantes. En effet, les autorités marocaines doivent concentrer leurs efforts sur les facteurs ci-dessus identifiés si elles souhaitent améliorer leur notation auprès de l'agence SP. Parmi ces facteurs, il y a ceux qui doivent jouir d'une certaine priorité vu leur importance dans le processus de notation de cette agence. Il s'agit, plus précisément, de la maîtrise du déficit budgétaire et de la réduction du poids de la dette totale du Trésor dans le revenu national.

5 Conclusion

Après avoir réalisé une analyse de la variable qu'on désire modéliser, on va expliquer dans les chapitre suivants les méthodes qu'on va choisir pour notre étude ainsi que les données et leurs sources.

Deuxième partie

Cadre théorique

Chapitre 2

Théorie de l'étude

1 Data science et Machine Learning

La science des données (Data Science) est le domaine d'étude qui combine l'expertise du domaine, les compétences en programmation et les connaissances en mathématiques et en statistiques pour extraire des informations significatives des données. Les praticiens de la science des données appliquent des algorithmes d'apprentissage automatique à des nombres, des textes, des images, des vidéos, des sons, et plus afin de produire des systèmes d'intelligence artificielle (IA) pour effectuer des tâches qui nécessitent habituellement l'intelligence humaine. À leur tour, ces systèmes génèrent des informations que les analystes et les utilisateurs professionnels peuvent traduire en valeur commerciale tangible.

Quant à l'apprentissage automatique (Machine Learning) est un sous-ensemble de l'intelligence artificielle (IA) dans lequel les algorithmes apprennent ,par l'exemple à partir de données historiques, pour prédire les résultats et découvrir des modèles difficilement repérables par les humains. Par exemple, l'apprentissage automatique peut révéler les clients qui sont susceptibles de se désabonner, les demandes d'assurance frauduleuses probables, etc. Bien que l'apprentissage automatique existe depuis les années 1950, les récentes percées dans les ressources de calcul à faible coût comme le stockage en nuage, la collecte de données plus facile et la prolifération de la science des données en ont fait la prochaine grande nouveauté dans l'analyse commerciale.

Le Machine Learning empreinte donc les techniques de Data Mining permettant de classer et trier les objets de notre monde (Classification), de prédire des événements (Prédiction), d'identifier des règles sous-jacentes à des données, d'analyser des séries temporelles , d'interpréter des textes (Analyse des sentiments sur les réseaux sociaux, NLP)... La figure

suivante donne un aperçu des classes de certains algorithmes d'apprentissage automatique dans le domaine.

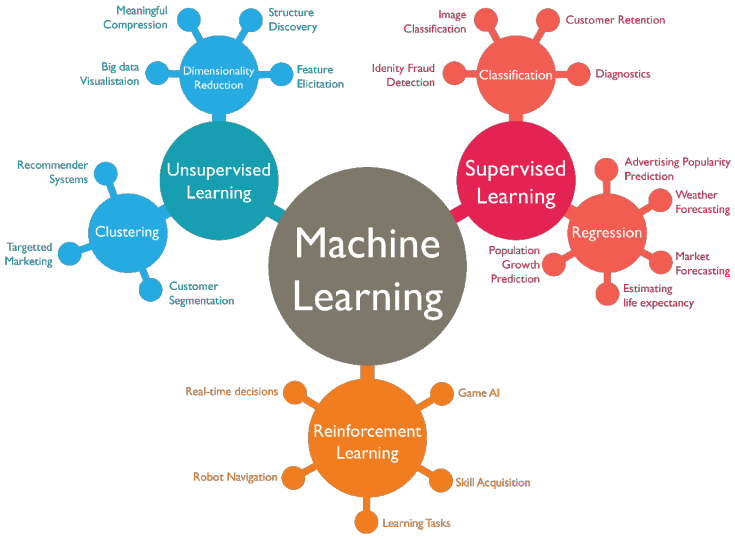

FIGURE 2.1 – Algorithmes d'apprentissage automatique

2 Machine learning supervisé

L'apprentissage automatique supervisé est une technologie élémentaire mais stricte. Les opérateurs présentent à l'ordinateur des exemples d'entrées et les sorties souhaitées, et l'ordinateur recherche des solutions pour obtenir ces sorties en fonction de ces entrées. Le but recherché est que l'ordinateur apprenne la règle générale qui mappe les entrées et les sorties.

Le machine learning avec supervision peut se subdiviser en deux types :

- **Classification :** La variable de sortie est une catégorie.
- **Régression :** La variable de sortie est une valeur spécifique.

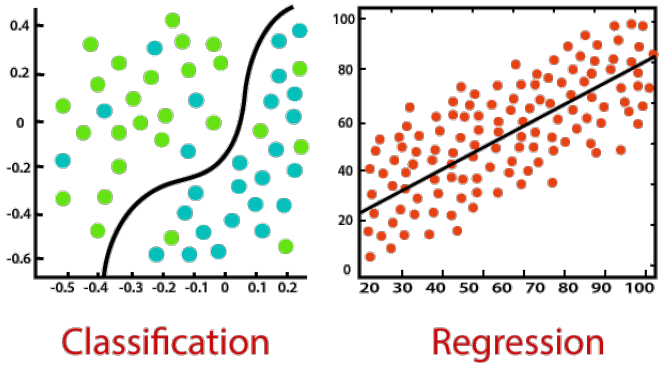

FIGURE 2.2 – Classification vs Regression

2.1 Régression

Afin d'estimer une valeur numérique en sortie, la régression s'avère efficace. Elle nous permet d'exhiber le lien entre les différentes variables dépendantes et indépendantes de notre base de données. Nous pourrions l'exploiter par exemple afin d'estimer la température (Variable quantitative) d'une zone précise en fonction de sa latitude et sa longitude. Dans ce même exemple, nous pourrions intégrer une variable qui décrit le temps, nous aurions affaire à une régression avec des données temporelles. La variable cible représenterait alors la température, le reste des variables représenterait les prédicteurs.
L'algorithme de régression choisi se chargera alors d'estimer la variable cible évaluée en fonction des prédicteurs. Cette étape comporte plusieurs itérations, les valeurs originales que nous avions utilisé pour la mise en place du modèle sont éliminées, et nous pourrons par conséquent appliquer ce modèle résultant sur des données différentes et avoir les résultats souhaités.
Il est important de souligner enfin que nous parlons de modèle de régression linéaire lorsque les paramètres du problème peuvent former une combinaison linéaire instantanément ou grâce à des transformations que nous pouvons entamer comme la technique de changement de variables. Par contre, il est impossible parfois, en fonction des données mises en jeu, d'obtenir un modèle linéaire.
Nous recourrons alors à des algorithmes différents appartenant, par exemple, aux estimations non paramétriques. Il nous faut enfin vérifier la qualité de notre régression et si elle est la plus convenable à nos données.

2.2 Classification

La classification est une méthode qui nous permet de prévoir un résultat, comme pour la régression, mais concerne principalement les variables cibles de type catégoriel. La génération du modèle s'effectue en fonction de variables numériques et qualitatives.
Il est nécessaire que nous ayons à notre disposition des données initiales décrivant les valeurs des prédicteurs et de la variable cible. Le processus de classification se charge alors de rechercher les liaisons entre les différents attributs qui vont nous permettre de trouver notre prévision. Une fois ceci établi, nous pourrions appliquer notre modèle de classification sur de nouvelles données totalement anonymes pour l'algorithme mais à condition qu'on ait les mêmes variables que ceux avec quoi on a construit notre modèle en ôtant la variable cible pendant ce processus. Une fois que l'algorithme est complété, la prévision est obtenue en fonction des données récentes.
L'étape finale est le jugement de la performance du modèle en question et sa validation à

travers des techniques adaptés à notre situation.

3 Machine learning Non supervisé

L'apprentissage non supervisé est une technique d'apprentissage automatique dans laquelle les utilisateurs n'ont pas besoin de superviser le modèle. Au lieu de cela, il permet au modèle de travailler par lui-même pour découvrir des modèles et des informations qui n'étaient pas détectés auparavant. Il traite principalement les données non étiquetées.

Les algorithmes d'apprentissage non supervisé permettent aux utilisateurs d'effectuer des tâches de traitement plus complexes que l'apprentissage supervisé. Cependant, l'apprentissage non supervisé peut être plus imprévisible par rapport aux autres méthodes d'apprentissage naturel.

Les algorithmes d'apprentissage non supervisé comprennent le regroupement et l'association.

3.1 Le regroupement

Le principe est assez simple : Nous regroupons les observations qui ont des degrés de similitude en un seul cluster homogène. Par conséquent, les observations appartenant à d'autres groupes auront des caractéristiques différentes.

Le regroupement (clustering) consiste à trouver la distribution sous-jacente des exemples dans leur espace de description. C'est-à-dire, à partir d'une base de données non étiquetées, nous cherchons à former des groupes (ou clusters) homogènes en fonction d'une certaine notion de similarité. Les observations qui sont considérées similaires sont associées au même groupe alors celles qui sont considérées comme différentes sont associées à des groupes différents.

Il existe plusieurs algorithmes capables d'effectuer le partitionnement des données, nous choisissons alors le meilleur par rapport aux attributs de l'ensemble des données à notre disposition.

3.2 L'association

Les règles d'association nous permettent d'établir des associations entre des objets de données dans de grandes bases de données. Cette technique non supervisée consiste à découvrir des relations intéressantes entre des variables dans de grandes bases de données. Par exemple, les personnes qui achètent une nouvelle maison sont plus susceptibles d'acheter de nouveaux meubles.

4 Apprentissage automatique semi-supervisé

L'apprentissage semi-supervisé est une approche d'apprentissage qui s'intéresse à l'étude de la façon dont les ordinateurs et les systèmes naturels tels que les humains apprennent en présence de données étiquetées et non étiquetées. Traditionnellement, l'apprentissage a été étudié soit dans l'approche non supervisée où toutes les données sont non étiquetées, soit dans l'approche supervisée où toutes les données sont étiquetées. L'objectif de l'apprentissage semi-supervisé est de comprendre comment la combinaison de données étiquetées et non étiquetées peut modifier le comportement d'apprentissage, et de concevoir des algorithmes qui tirent parti d'une telle combinaison.

5 Conclusion

Dans ce projet on est censé établir la modélisation de la variable **"Notation Souveraine"**. Ceci dit, notre variable cible est une variable qualitative. Alors nous procéderons dans le chapitre suivant aux différentes méthodes de classification qu'on peut utiliser.

Chapitre 3

Revue empirique

1 La méthode de la régression logistique

La classification peut être effectuée grâce à la méthode de régression logistique. Elle appartient aux techniques d'analyse en présence de plusieurs variables. Son importance réside dans l'estimation de la liaison entre l'émergence d'un phénomène qui sera de type qualitatif et les éléments qui pourraient potentiellement agir sur lui. En d'autres termes, ces éléments représentent les variables explicatives du problème.

1.1 Régression logistique binaire

Par rapport à notre domaine d'étude qui est le secteur financier, cette méthode de régression logistique binaire est valable pour prédire la catégorie à laquelle notre variable cible "Notation souveraine" appartient :

— **Notes d'investissement :** AAA, AA, A, BBB.

— **Notes Spéculatives :** BB, B, CCC, CC, C, D.

Afin de mieux comprendre le fonctionnement pratique de cette méthode, essayons d'abord d'assimiler son aspect théorique :

Supposons qu'on dispose de la probabilité estimée (notée Y) qu'un commentaire est positif. Nous posons $u = C_0 + C_1X_1 + C_2X_2 + ... + C_kX_k$ qui décrit la régression linéaire effectuée avec $C_0, C_1, C_2, ..., C_k$ qui représentent des valeurs constantes. L'ensemble des variables $X_1, X_2, ..., X_k$ constituent les variables indépendantes du problème. La définition de Y s'écrit

de la façon suivante :

$$Y = \frac{e^u}{1 + e^u}$$

La fonction logistique est une courbe en forme de S capable de convertir tout nombre réel en une valeur comprise entre 0 et 1, sans toucher ces limites. Nous voyons ci-dessous un tracé (figure[3.1]) qui transforme les nombres entre −8 et 8 en une gamme de 0 à 1 en utilisant la fonction *sigmoïde* ou *logistique*.

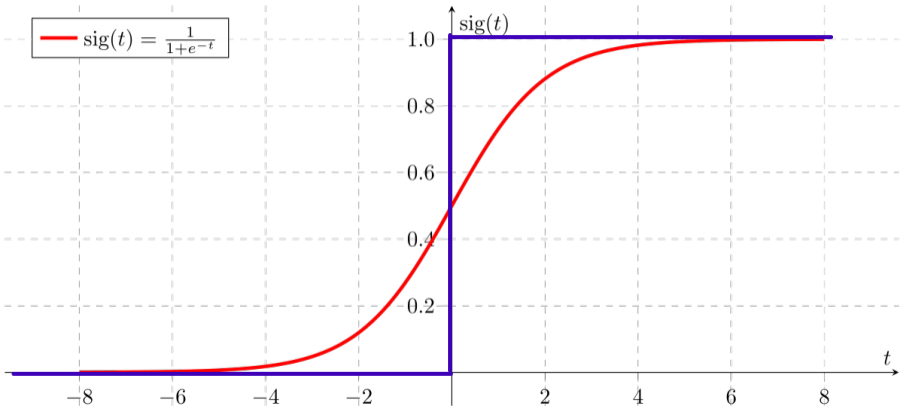

FIGURE 3.1 – Exemple de transformation de la fonction logistique

1.2 Régression logistique multinomiale

La régression logistique peut être étendue pour traiter les réponses qui sont polytomiques, c'est-à-dire qui prennent r $>$ 2 catégories. Lors de l'analyse d'une réponse polytomique, il est important de noter si la réponse est ordinale (constituée de catégories ordonnées) ou nominale (constituée de catégories non ordonnées). Certains types de modèles ne sont appropriés que pour les réponses ordinales ; d'autres modèles peuvent être utilisés que la réponse soit ordinale ou nominale. Si la réponse est ordinale, nous ne devons pas nécessairement tenir compte de l'ordre, mais il est souvent utile de le faire. L'utilisation de l'ordre naturel peut :

— Conduire à un modèle plus simple et plus parcimonieux.

— Augmenter la capacité à détecter les relations avec d'autres variables.

Dans le cas présent on va travailler sur la notation souveraine ayant plusieurs modalités ordonnées. Donc notre modèle est ordonné.

Dans un modèle ordonné, les modalités de la variable à expliquer sont hiérarchisées. Elles indiquent l'appartenance de l'individu à une classe ou à une catégorie, par exemple l'appartenance à une tranche de revenu. Nous pouvons distinguer deux classes de modèles à choix multiples ordonnés en fonction de la variable à expliquer qui est issue, soit d'une « discrétisation » d'une variable continue telle que l'appartenance à une tranche de salaire, soit directement d'une appartenance à une catégorie (faire du sport : une fois par semaine, une fois par mois, une fois par an, ...).
Nous modélisons une variable latente continue :

$$y_i^* = a_0 + a_1 x_i + \epsilon_i$$

Les valeurs prises par la variable y_i correspondent à des intervalles dans lesquels se trouve y_i^* définissant ainsi le modèle de décision suivant à M + 1 modalités :

$$\begin{cases} y_i = 0 & \text{si} \quad y_i^* \leqslant c_1 \\ y_i = 1 & \text{si} \quad c_1 \leq y_i^* \leqslant c_2 \\ y_i = 2 & \text{si} \quad c_2 \leq y_i^* \leqslant c_3 \\ \vdots = \vdots & \\ y_i = M & \text{si} \quad c_M \leq y_i^* \end{cases} \tag{3.1}$$

Soit P_i la probabilité d'apparition de chaque événement pour l'individu i :

$$\begin{aligned} P_{i0} = Prob(y_i = 0) &= \Phi(c_1 - (a_0 + a_1 x_i)) \\ P_{i1} = Prob(y_i = 1) &= \Phi(c_2 - (a_0 + a_1 x_i)) - \Phi(c_1 - (a_0 + a_1 x_i)) \\ P_{i2} = Prob(y_i = 2) &= \Phi(c_3 - (a_0 + a_1 x_i)) - \Phi(c_2 - (a_0 + a_1 x_i)) \\ &\dots \\ P_{iM} = Prob(y_i = M) &= 1 - \Phi(c_M - (a_0 + a_1 x_i)) \end{aligned}$$

Avec Φ la fonction de répartition de la loi de probabilité normale ou logistique $(\Phi(t) = \frac{e_t}{1+e_t})$ et $\sum_{i=0}^{M} P_i = 1$

Le recours à une fonction de répartition normale, permet de définir un modèle de type Probit et une fonction de répartition de type logistique permet de définir un modèle Logit. L'estimation de tous les paramètres, les coefficients de régression (a_i) et les valeurs des seuils (c_i) des modèles ordonnés (Probit ou Logit) est effectuée à l'aide des algorithmes de maximisation d'une fonction de Log-vraisemblance définie par les P_{ij}.

Les valeurs des coefficients des modèles ne sont pas directement interprétables en termes de propension marginale, seuls les signes des coefficients indiquent si la variable agit positivement ou négativement sur la variable latente.

Les résultats d'estimation s'apprécient de la même manière que pour les modèles de choix binaires :

— la significativité des coefficients à l'aide des ratios z-Statistique,

— la significativité globale de l'ajustement (l'hypothèse : $H_0 : a_1 = a_2 = a_3 = \ldots = a_k = 0$) par la statistique $LR = -2(Ln(L_R) - Ln(L_u))$ qui suit sous l'hypothèse nulle H_0, une distribution d'un χ^2 à k degrés de liberté.
Le $pseudo - R^2$ est donné par : $R^2 = 1 - \frac{Log(L_u)}{Log(L_R}$

2 La méthode machine à vecteurs support (SVM)

Le SVM est un algorithme d'apprentissage automatique supervisé qui aide à résoudre les problèmes de classification ou de régression. Il vise à trouver une frontière optimale entre les sorties possibles.

En termes simples, le SVM effectue des transformations complexes des données en fonction du noyau sélectionné et, sur la base de ces transformations, il tente de maximiser les limites de séparation entre les points de données en fonction des étiquettes ou des classes que nous avons définies.

Dans sa forme de base, la séparation linéaire, le SVM essaie de trouver une ligne qui maximise la séparation entre un ensemble de données à deux classes de points dans un espace à deux dimensions. Pour généraliser, l'objectif est de trouver un hyperplan qui maximise la séparation des points de données vers leurs classes potentielles dans un espace à n dimensions. Les points de données présentant la distance minimale à l'hyperplan (points les plus proches) sont appelés vecteurs de support.

3 La méthode Random Forest

Une grande partie du Machine Learning est la classification - nous voulons savoir à quelle classe (ou groupe) appartient une observation. La capacité de classer précisément les observa- tions est extrêmement précieuse pour diverses applications commerciales, comme la prévision de l'achat d'un produit par un utilisateur particulier ou la prévision de la défaillance ou non d'un prêt donné.

Le Random Forest est un algorithme d'apprentissage supervisé qui fait partie des méthodes d'ensembles. Le principe des méthodes d'ensemble est basé sur le regroupement de plusieurs algorithmes d'apprentissage instables pour créer un algorithme plus performant.

Le Random Forest est le regroupement de plusieurs arbres de décision, où chaque arbre est construit sur un échantillon tiré avec remise de la base de données d'apprentissage, avec un choix aléatoire des variables de séparation. La prédiction du modèle est la moyenne des prédictions de l'ensemble des arbres de décisions dans le cas de la régression, et le vote majoritaire dans le cas de la classification.

La performance du Random Forest vient du regroupement de plusieurs arbres de décision, qui permet de réduire la variabilité dans la variable sortie, et le choix aléatoire des variables de séparation qui permet de décorréler les arbres de décision construites, et par conséquent réduire l'erreur final.

Troisième partie

Cadre pratique

Chapitre 4

Mise en oeuvre du système

1 Outils choisis

1.1 Python

Selon des études récentes, Python est le langage de programmation préféré des scientifiques de données. Ils ont besoin d'un langage facile à utiliser, avec une disponibilité de bibliothèque décente et une grande participation de la communauté. Les projets qui ont des communautés inactives sont généralement moins susceptibles de main- tenir ou de mettre à jour leurs plateformes, ce qui n'est pas le cas de Python.
Qu'est-ce qui rend Python si idéal pour la science des données ? Nous avons examiné pourquoi Python est si répandu dans l'industrie florissante de la science des données - et comment vous pouvez l'utiliser pour vos grands projets de données et d'apprentissage de la machine.

Python est connu depuis longtemps comme un langage de programmation simple à reprendre, du point de vue de la syntaxe en tout cas. Python a également une communauté active avec un vaste choix de bibliothèques et de ressources. Le résultat ? Nous disposons d'une plate-forme de programmation qu'il est logique d'utiliser avec les technologies émergentes comme l'apprentissage machine et la science des données.

Les professionnels qui travaillent avec des applications en sciences des données ne veulent pas s'embourber dans des exigences de programmation compliquées. Ils veulent utiliser des langages de programmation comme Python pour effectuer des tâches sans problèmes.

La science des données consiste à extrapoler des informations utiles à partir d'énormes réserves de statistiques, de registres et de données. Ces données sont généralement non triées

et difficiles à corréler avec une précision significative. L'apprentissage machine peut établir des connexions entre des ensembles de données disparates, mais il nécessite une grande puissance de calcul.

Python répond à ce besoin en étant un langage de programmation polyvalent. Il permet de créer des sorties CSV pour faciliter la lecture des données dans un tableur. Il est également possible d'obtenir des sorties de fichiers plus complexes qui peuvent être ingérées par des groupes de Machine Learning pour le calcul.

FIGURE 4.1 – Logo de Python

1.2 Bibliothèque Pandas

Pandas est une bibliothèque Python qui est un outil simple mais puissant de la science des données. Python Pandas est l'un des paquets les plus utilisés en Python. Ce paquet comprend de nombreuses structures de données et des outils pour manipuler et analyser efficacement les données. Python Pandas est utilisé dans des domaines tels que l'économie, la comptabilité, l'analyse, les statistiques, etc. partout, y compris dans les secteurs commerciaux et universitaires.

Nous avons choisi d'utiliser la bibliothèque Pandas en Python pour suivre, évaluer et purifier les données. Python Pandas est bien adapté aux données de différentes formes, telles que :

— Les données tabulaires dont les colonnes sont tapées de manière hétérogène.

— Les données sur les séries chronologiques ordonnées et non ordonnées.

— Les données matricielles.

— Les données non identifiées.

— Toute autre forme d'ensemble de données statistiques ou d'observations.

FIGURE 4.2 – Logo de Pandas

1.3 Bibliothèque Scikit-learn

Scikit-learn est une bibliothèque du « Machine Learning » qui existe dans « Python » gratuitement. Elle propose divers algorithmes de classification, de régression et de regroupement, notamment des machines à vecteurs de support, Random Forest, ect. . .

Elle est conçue pour interagir avec les bibliothèques numériques et scientifiques de Python, tel que : NumPy et SciPy.

FIGURE 4.3 – Logo de Scikit-learn

1.4 Google Colab

Google Colaboratory, parfois appelé Colaboratory ou google colab en abrégé, est un service basé sur le python fourni par google avec un concept similaire à Jupyter Notebook qui fonctionne dans le nuage, aucune installation requise pour l'utiliser les utilisateurs peuvent y accéder gratuitement à partir du navigateur, avec un compte google.

Google Colaboratory fournit aux utilisateurs un environnement python complet plus des tonnes de bibliothèques d'apprentissage automatique avec une RAM 2 , un CPU 3 , un GPU 4 et même un TPU 5 gratuits pour accélérer votre travail, l'écriture du code se fait à travers un paradigme orienté cellule, il permet également de combiner le code et le texte riche dans un seul document, ainsi que des images, HTML, LaTeX et plus encore, lors de la création de carnets colab, ils sont enregistrés automatiquement dans le lecteur google de l'utilisateur, ce qui facilite le partage entre les développeurs.

FIGURE 4.4 – Logo de Google Colab

2 Présentation des données

Avant d'exposer les différents composants de notre base de données, préparons d'abord l'environnement où nous allons travailler. Afin de mieux structurer notre tâche de programmation, nous avons choisi d'utiliser Google Colab et le langage de programmation Python pour coder.

Ceci aussitôt établi, nous devons désormais importer les différentes bibliothèques que nous allons exploiter. Ces bibliothèques sont :

- Pandas : Pour le traitement des bases de données.
- Numpy : pour l'insertion des grands tableaux et matrices multidimensionnels, ainsi qu'une vaste collection de fonctions mathématiques de haut niveau pour opérer sur ces tableaux.

La phase de création de la base de données est l'une des étapes primordiales d'un projet, la fiabilité des résultats dépend essentiellement de la qualité des données. Les données qu'on a utilisé dans cette modélisation sont extraites du site officiel de l'agence Standard & Poor's. Ceci dit a, on a réussi à regrouper 51 variables explicatives qui peuvent être classifiées selon 6 groupements de données :

- Données économiques
- Données monétaires
- Données gouvernementales
- Données de la balance des paiements
- Données du bilan externe
- Données relatives aux dettes et emprunts du gouvernement

Le premier groupement comporte 10 variables :

- **Nominal_GDP(bil.LC)** : c'est le produit intérieur brut (PIB) nominal sans tenir compte des variations de prix et de l'inflation (en monnaie locale).
- **Nominal_GDP(bil.US$)** : c'est le produit intérieur brut (PIB) nominal sans tenir compte des variations de prix et de l'inflation (en Dollars).
- **GDP_per_capita(US$)** : ou le PIB par habitant, c'est la valeur du PIB divisée par le nombre d'habitants d'un pays. Il est plus efficace que le PIB pour mesurer le développement d'un pays.
- **Real_GDP_growth(%)** : le taux de croissance du PIB réel, il mesure l'évolution d'une période à l'autre du PIB réel. Le PIB réel est le produit intérieur brut qui tient compte de l'inflation. Cet indicateur est exprimé en pourcentage.
- **Real_GDP_per_capita_growth(%)** : Cet indicateur mesure le taux de croissance de la variable PIB réel par habitant qui a déjà été expliqué précédemment.
- **Real_investment_growth(%)** : il s'agit du taux de croissance de l'investissement matériel, en effet l'investissement matériel est l'ensemble d'argents qui a été investi dans des équipements, des machines ....
- **Investment/GDP(%)** : c'est un indicateur qui mesure le rapport entre l'investissement et le PIB, il est exprimé en pourcentage. Il calcule la part du PIB qui a été investi.
- **Savings/GDP(%)** : C'est le rapport entre l'épargne et le PIB, il mesure la part du PIB qui a été épargné.
- **Exports/GDP(%)** : Le rapport des exportations sur le PIB, il exprime la part du PIB qui a été consacré aux exportations.
- **Unemployment_rate(%of_workforce)** : C'est le taux de chômage, il exprime le pourcentage de chômeurs dans la population active (actifs occupés + chômeurs).

Le deuxième groupement de variables concerne les données relatives à la politique monétaire, et on y trouve :

- **CPI_growth(%)** : c'est le taux de croissance de l'indice des prix à la consommation (IPC). L'IPC est un instrument de mesure de l'inflation. Il permet d'estimer, entre deux périodes données, la variation moyenne des prix des produits consommés par les ménages.

— **GDP_deflator_growth(%)** : ou le taux de croissance du déflateur du PIB, il est à son tour un indicateur qui mesure l'inflation et permet de corriger les agrégats des effets de l'inflation.

— **Exchange_rate_year-end(LC/$)** : Taux de change de fin d'année.

— **Banks_claims_on_resident_non-govt_sector_growth** : il s'agit du taux de croissance des créances bancaires sur les non-gouvernementaux résidents.

— **Banks_claims_on_resident_non-govt_sector/GDP** : c'est le rapport entre les créances bancaires sur les non-gouvernementaux résidents et le PIB.

— **Foreign_currency_share_of_claims_by_banks_on_residents** : il s'agit de la part en monnaie étrangère des créances bancaires sur les résidents.

— **Foreign_currency_share_of_residents_bank_deposits** : il s'agit de la part des dépôts bancaires des résidents en devises étrangères.

— **Real_effective_exchange_rate_growth** : C'est la croissance du taux de change effectif réel (TCER) ; le TCER d'une monnaie est défini comme le taux de change effectif nominal rapporté aux prix relatifs entre le pays considéré et ceux des principaux pays partenaires et concurrents.

On trouve par la suite les données gouvernementales avec 10 variables différentes :

— **GG_balance/GDP(%)** : C'est le rapport du solde budgétaire des administrations publiques sur le PIB, et il est exprimé en pourcentage. Le solde budgétaire des administrations publiques est la différence entre leurs recettes et leurs dépenses.

— **Change_in_Net_GG_debt/GDP(%)** : il s'agit du rapport entre la variation de la dette nette des administrations publiques et le PIB.

— **Primary_GG_balance/GDP(%)** : c'est le rapport entre le solde budgétaire primaire des administrations publiques sur le PIB, on désigne par solde primaire la situation budgétaire hors paiement des intérêts. Cet indicateur est utilisé pour connaître le solde budgétaire permettant de stabiliser ou de diminuer l'endettement.

— **GG_Revenues/GDP(%)** : il s'agit du rapport entre les recettes publiques et le PIB. Ces recettes constituent l'ensemble des impôts, des taxes et cotisations sociales.

— **GG_Expenditures/GDP(%)** : c'est le rapport entre les dépenses publiques et le PIB, cet indicateur est exprimé en pourcentage.

— **GG_interest_expenditure/revenues(%)** : il s'agit du rapport des charges d'intérêt des administrations publiques sur les recettes publiques.

— **Gross_GG_debt/GDP(%)** : il s'agit de la valeur de la dette brute des administrations publiques sur le PIB.

— **Debt/Revenues(%)** : le rapport entre la dette et les recettes des administrations publiques.

— **Net_GG_debt/GDP(%)** : le rapport entre la dette nette des administrations publiques et le PIB.

— **Liquid_assets/GDP** : c'est le rapport entre les actifs liquides et le PIB ; un actif liquide est un actif qui peut être facilement converti en espèces.

Le quatrième groupement concerne les données relatives à la balance de paiement, il s'agit de l'ensemble des données des échanges de biens, services et de capitaux pendant une période donnée entre les agents économiques résidents d'un pays et le reste du monde :

— **CARs/GDP(%)** :(CARs : Current account receipts) il s'agit du rapport entre les recettes du compte courant et le PIB, cet indicateur est exprimé en pourcentage. En effet, le compte courant d'un pays enregistre la valeur des exportations et des importations de biens et de services et les transferts internationaux de capitaux.

— **Real_exports_growth(%)** : il s'agit du taux de croissance des exportations réelles.

— **Current_account_balance/GDP(%)** : C'est le rapport entre le solde du compte courant et le PIB, il est exprimé en pourcentage.

— **Current_account_balance/CARs(%)** : C'est le rapport entre le solde du compte courant et ses recettes, exprimé en pourcentage.

— **Usable_reserves/CAPs(months)** : c'est le rapport entre les réserves utilisables et les dépenses du compte courant. (CAPs : Current account payments)

— **Gross_ext.fin.needs/(CAR+use.res.)(%)** : c'est le rapport entre le besoin de financement extérieur brut et les recettes du compte courant.

— **Net_FDI/GDP(%)** : il s'agit du rapport entre les investissements directs étrangers nets et le PIB. Les investissements directs à l'étranger sont les investissements réalisés par une entreprise en direction d'une entreprise étrangère.

— **Trade_balance/GDP** :C'est la balance commerciale sur le PIB, en notant que la balance commerciale est la différence entre les exportations et les importations des biens et services.

— **Net_portfolio_equity_inflow/GDP** : C'est le rapport du portefeuille des fonds nets entrants sur le PIB.

Par la suite, on trouve les données relatives au bilan externe avec 5 variables :

- **Narrow_net_ext.debt/CARs(%)** : il s'agit du rapport entre la dette extérieure nette au sens étroit, qui est égale à la dette extérieure totale moins les liquidités des actifs extérieures, et les recettes du compte courant.
- **Narrow_net_ext.debt/CAPs(%)** : c'est le rapport entre la dette extérieure nette au sens étroit et les dépenses du compte courant.
- **Net_ext.liabilities/CARs(%)** : c'est le rapport entre les passifs extérieurs nets et les recettes du compte courant.
- **Short-term_external_debt_by_remaining_maturity/CARs** : C'est le rapport entre la dette extérieure à court terme et les recettes du compte courant.
- **Usable_reserves(US$mil.)** : : Il s'agit des réserves utilisables.

Le dernier groupement concerne les données relatives aux dettes et emprunts du gouvernement :

- **Gross_LT_commercial_borrowing(US$bil.)** :il s'agit des emprunts commerciaux bruts à long terme.
- **Commercial_debt_stock(year_end_US$bil.)** : il s'agit de l'encours de la dette commerciale, en supposant que la dette commerciale est une dette qui découle d'un acte de commerce.
- **ST_debt(US$bil.)** : cet indicateur signifie la dette à court terme, c'est-à-dire une dette d'une période de moins d'un an.
- **Bi-/Multilateral_debt(%of_total)** : il s'agit des dettes publiques bilatérales et multilatérales. En effet une dette bilatérale est l'ensemble des emprunts contractés par un Etat auprès d'un autre Etat. Par ailleurs, la dette multilatérale signifie les prêts contractés auprès des institutions financières internationales comme la Banque mondiale, le Fonds Monétaire International (FMI)... Cet indicateur représente la part des dettes bilatérales et multilatérales du total des dettes.
- **ST_debt(%of_total)** : c'est le pourcentage des dettes à court terme du total des dettes.
- **FC_debt(%of_total)** : c'est le pourcentage des dettes en devises du total des dettes.
- **LT_fixed-rate_debt(%of_total)** : : il s'agit de la part des dettes à taux fixe à long terme du total ; en effet le taux d'intérêt est fixé lors de l'émission de l'obligation et reste le même durant toute la vie de l'obligation.

— **Roll-over ratio (% of debt)** :c'est le pourcentage du taux de refinancement de la dette ; l'opération de refinancement de la dette consiste au remboursement d'un emprunt auprès d'un établissement de crédit suivi de la souscription d'un nouvel emprunt.

— **Roll-over_ratio(%of_GDP)** : c'est le pourcentage du taux de refinancement de la dette du PIB.

Notre base de données contient 931 lignes et 54 variables. Elle concerne les indicateurs sur 133 pays sur une période de 7 ans de 2015 à 2021.

3 Préparation des données

Afin de pouvoir générer et appliquer un modèle de Machine Learning, l'une des étapes les plus cruciales est la préparation des données afin de pouvoir les adapter au modèle choisi. Ceci dit, nous devons désormais sélectionner les variables qui auront un poids significatif pour l'accomplissement de notre prévision.

En premier lieu, avant d'importer notre base de données sur Python, on a remplacé quelques valeurs manquantes par la médiane des autres valeurs présentes par pays. Et on a supprimé 3 variables appartenant au groupement des données monétaires vu qu'elle contiennent un pourcentage élevé de valeurs manquantes, ce sont :

- **Foreign_currency_share_of_claims_by_banks_on_residents**.
- **Foreign_currency_share_of_residents_bank_deposits**.
- **Real_effective_exchange_rate_growth**.

Et on supprime également les deux variables **Year** et **Country** vu que ces deux variables ne vont pas servir dans notre modélisation.
En deuxième lieu, on importe notre base de données sur Python pour avoir une idée claire sur ses composantes. Le nombre d'observations et de variables a diminué :

```
[5]  print("Notre base de données contient {} lignes et {} colonnes.".format(df.shape[0], df.shape[1]))

     Notre base de données contient 881 lignes et 49 colonnes.
```

FIGURE 4.5 – La forme de notre BD

4 Nettoyage de données

Le nettoyage des données représente une partie importante dans la compréhension des données et dans l'assurance d'une qualité de prévision supérieure. Dans cette étape, nous nous chargeons de repérer et remanier ou ôter les observations qui souffrent d'erreurs de saisie ou de problèmes de redondance. Si jamais nous songions à éviter d'appliquer ce processus, nous aurions des estimations erronées avec des prédictions déplorables.

C'est à vrai dire l'origine des données qui est responsable de sa qualité, l'adaptation et le nettoyage de ces données est l'une des tâches d'un Data-Scientist. Nous connaissons plusieurs erreurs possibles, nous pouvons citer :

- Les erreurs de mesure
- Les erreurs de saisie de données
- Les possibilités de redondance

Le processus de nettoyage de données se fait comme suit :

4.1 Valeurs manquantes

Après avoir recherché les valeurs manquantes, on remarque qu'il les pourcentages de ces valeurs par variable ne sont pas élevées. alors on décide de les éliminer vu qu'on n'a pas par quoi les remplacer.

```
Rating                                                  0.000000
Liquid_assets/GDP                                       0.000000
Debt/Revenues(%)                                        0.000000
Gross_GG_debt/GDP(%)                                    0.000000
GG_Expenditures/GDP(%)                                  0.000000
GG_Revenues/GDP(%)                                      0.000000
Primary_GG_balance/GDP(%)                               0.000000
Change_in_Net_GG_debt/GDP(%)                            0.000000
GG_balance/GDP(%)                                       0.000000
Exchange_rate_year-end(LC/$)                            0.000000
GDP_deflator_growth(%)                                  0.000000
Net_GG_debt/GDP(%)                                      0.000000
Nominal_GDP(bil.LC)                                     0.000000
CPI_growth(%)                                           0.000000
Nominal_GDP(bil.US$)                                    0.000000
GDP_per_capita(US$)                                     0.000000
Real_GDP_per_capita_growth(%)                           0.000000
Real_GDP_growth(%)                                      0.000000
Commercial_debt_stock(year_end_US$bil.)                 0.007946
Gross_LT_commercial_borrowing(US$bil.)                  0.007946
Roll-over ratio (% of debt)                             0.007946
Roll-over_ratio(%of_GDP)                                0.007946
ST_debt(%of_total)                                      0.007946
FC_debt(%of_total)                                      0.007946
LT_fixed-rate_debt(%of_total)                           0.007946
GG_interest_expenditure/revenues(%)                     0.007946
Bi-/Multilateral_debt(%of_total)                        0.009081
ST_debt(US$bil.)                                        0.009081
Narrow_net_ext.debt/CAPs(%)                             0.015891
Narrow_net_ext.debt/CARs(%)                             0.015891
Exports/GDP(%)                                          0.015891
Trade_balance/GDP                                       0.015891
Banks_claims_on_resident_non-govt_sector_growth         0.015891
Current_account_balance/CARs(%)                         0.015891
Current_account_balance/GDP(%)                          0.015891
Banks_claims_on_resident_non-govt_sector/GDP            0.015891
Gross_ext.fin.needs/(CAR+use.res.)(%)                   0.015891
Short-term_external_debt_by_remaining_maturity/CARs     0.015891
CARs/GDP(%)                                             0.019296
Usable_reserves/CAPs(months)                            0.023837
Net_ext.liabilities/CARs(%)                             0.023837
Usable_reserves(US$mil.)                                0.023837
Net_portfolio_equity_inflow/GDP                         0.031782
Net_FDI/GDP(%)                                          0.031782
Investment/GDP(%)                                       0.031782
Savings/GDP(%)                                          0.031782
Real_exports_growth(%)                                  0.036322
Real_investment_growth(%)                               0.047673
Unemployment_rate(%of_workforce)                        0.087401
dtype: float64
```

FIGURE 4.6 – Pourcentage des valeurs manquantes

Après les avoir supprimé, la forme de notre base de données changera en 732 lignes et 49 colonnes.

```
Rating                                                  0.0
CARs/GDP(%)                                             0.0
Real_exports_growth(%)                                  0.0
Current_account_balance/GDP(%)                          0.0
Current_account_balance/CARs(%)                         0.0
Usable_reserves/CAPs(months)                            0.0
Gross_ext.fin.needs/(CAR+use.res.)(%)                   0.0
Net_FDI/GDP(%)                                          0.0
Trade_balance/GDP                                       0.0
Net_portfolio_equity_inflow/GDP                         0.0
Narrow_net_ext.debt/CARs(%)                             0.0
Narrow_net_ext.debt/CAPs(%)                             0.0
Net_ext.liabilities/CARs(%)                             0.0
Short-term_external_debt_by_remaining_maturity/CARs     0.0
Usable_reserves(US$mil.)                                0.0
Gross_LT_commercial_borrowing(US$bil.)                  0.0
Commercial_debt_stock(year_end_US$bil.)                 0.0
ST_debt(US$bil.)                                        0.0
Bi-/Multilateral_debt(%of_total)                        0.0
ST_debt(%of_total)                                      0.0
FC_debt(%of_total)                                      0.0
LT_fixed-rate_debt(%of_total)                           0.0
Liquid_assets/GDP                                       0.0
Roll-over ratio (% of debt)                             0.0
Net_GG_debt/GDP(%)                                      0.0
Gross_GG_debt/GDP(%)                                    0.0
Nominal_GDP(bil.LC)                                     0.0
Nominal_GDP(bil.US$)                                    0.0
GDP_per_capita(US$)                                     0.0
Real_GDP_growth(%)                                      0.0
Real_GDP_per_capita_growth(%)                           0.0
Real_investment_growth(%)                               0.0
Investment/GDP(%)                                       0.0
Savings/GDP(%)                                          0.0
Exports/GDP(%)                                          0.0
Unemployment_rate(%of_workforce)                        0.0
CPI_growth(%)                                           0.0
GDP_deflator_growth(%)                                  0.0
Exchange_rate_year-end(LC/$)                            0.0
Banks_claims_on_resident_non-govt_sector_growth         0.0
Banks_claims_on_resident_non-govt_sector/GDP            0.0
GG_balance/GDP(%)                                       0.0
Change_in_Net_GG_debt/GDP(%)                            0.0
Primary_GG_balance/GDP(%)                               0.0
GG_Revenues/GDP(%)                                      0.0
GG_Expenditures/GDP(%)                                  0.0
GG_interest_expenditure/revenues(%)                     0.0
Debt/Revenues(%)                                        0.0
Roll-over_ratio(%of_GDP)                                0.0
dtype: float64
```

FIGURE 4.7 – Élimination des valeurs manquantes

4.2 Les doublons

On remarque que notre base de données ne contient pas de doublons. La collecte de nos données était bien ciblée c'est pour cette raison qu'on ne trouve pas d'informations redondantes.

```
[8]  print("Nous avons  {} doublons dans Df.".format(df.duplicated().sum()))

     Nous avons  0 doublons dans Df.
```

FIGURE 4.8 – Le nombre de doublons

4.3 Valeurs aberrantes

Lors de la modélisation, il est important de nettoyer l'échantillon de données pour s'assurer que les observations représentent au mieux le problème.

Parfois, un ensemble de données peut contenir des valeurs extrêmes qui se situent en dehors de la fourchette attendue et ne ressemblent pas aux autres données. Ces valeurs sont appelées valeurs aberrantes et, souvent, la modélisation de l'apprentissage automatique et les compétences en matière de modélisation en général peuvent être améliorées en comprenant et même en supprimant ces valeurs aberrantes.

Dans notre cas, la nature de notre base de données ne nous permet pas d'éliminer les valeurs aberrantes. Les indicateurs de la notation souveraine diffèrent d'un pays à l'autre, en effet les pays du monde se répartissent en plusieurs catégories.

5 Analyse exploratoire des données

5.1 Variable cible

La variable qu'on désire modéliser est la notation souveraine établie par l'agence **Standard & Poor's**. On visualise en premier lieu les modalités de cette variable :

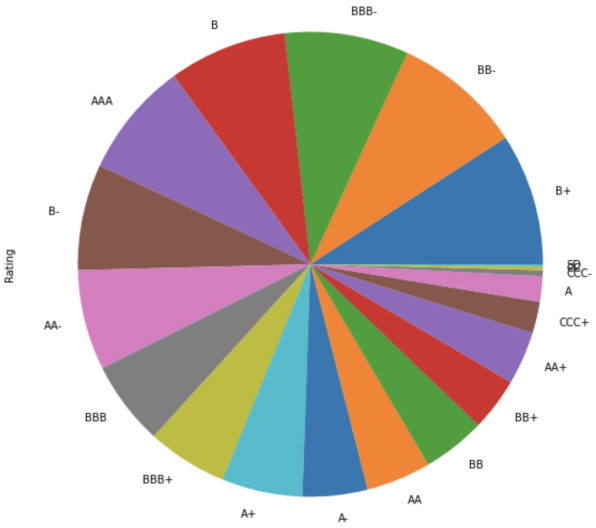

FIGURE 4.9 – Les modalités initiales de la variable **Rating**

D'après la figure [4.9], on remarque qu'on a 17 modalités, et on a des modalités dont on ne dispose pas d'observations suffisantes pour entraîner nos modèles.

Alors, on a établi quelques modifications sur notre base de données en réduisant le nombre de modalités tout en l'adaptant à notre étude.

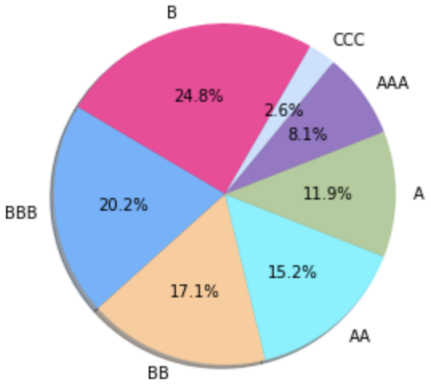

FIGURE 4.10 – Les modalités de la variable **Rating** après modification

5.2 Corrélation entre les variables explicatives

Dans cette partie, on essaie d'étudier les liaisons entre les variables explicatives. On remarque d'après le Heatmap ci-dessous qu'il y a plusieurs variables ayant une forte corrélation entre elles.

Quand deux variables explicatives sont fortement corrélées, elles donnent la même information en modélisation et elles risquent d'influencer notre modèle négativement. On va essayer de remédier à ce problème dans le chapitre prochain.

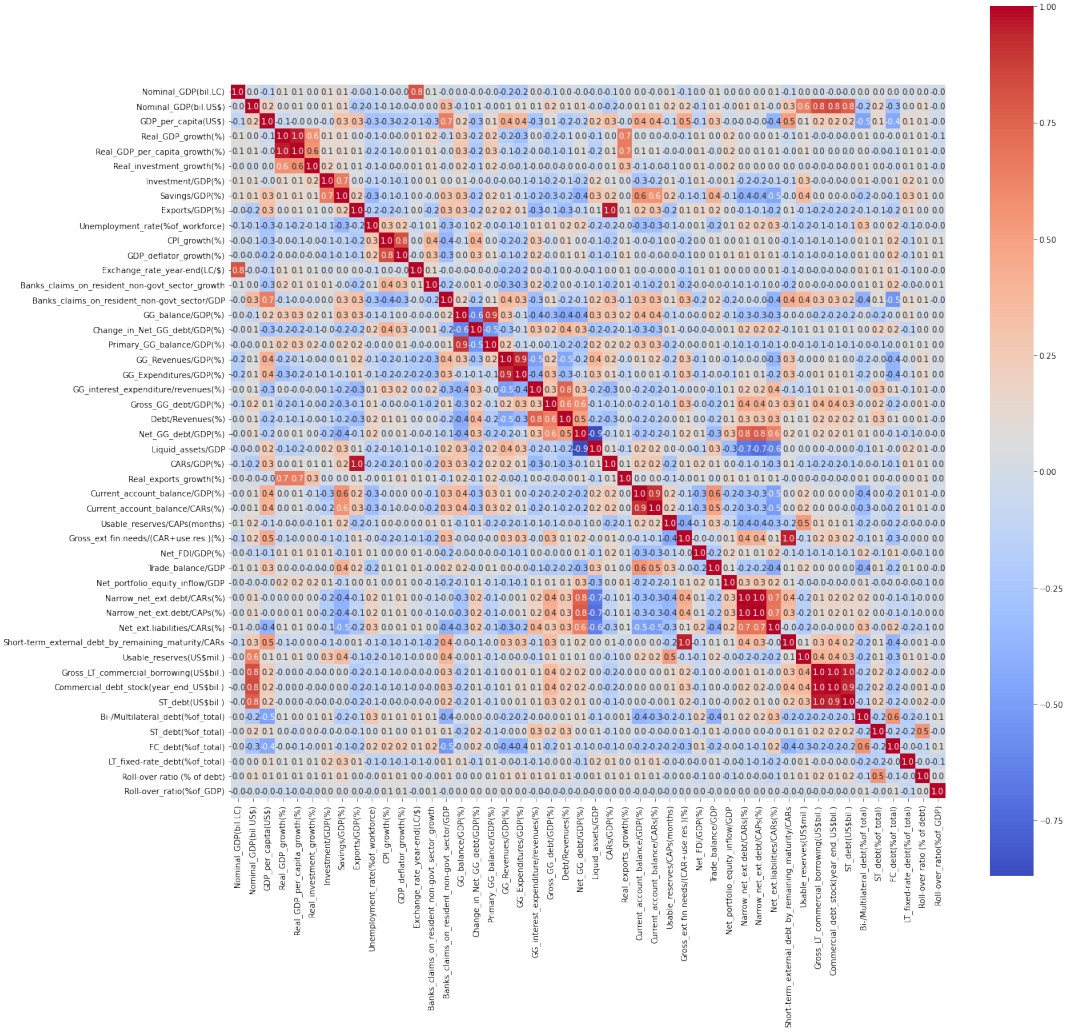

FIGURE 4.11 – Heatmap visualisant la corrélation entre les variables

Chapitre 5

Modélisations prédictives

Le processus de sélection est une étape importante dans la réalisation des modèles prédictifs. En effet, lorsque ce processus est accompli, on représente les données sous forme de matrice de telle façon à ce que les colonnes représentent les variables et les lignes qui décrivent les observations (ou individus). Ces variables sont ensuite fusionnées de façon à ce qu'elle forme une seule base de données de 40 colonnes qu'on utilise ensuite pour produire les algorithmes que nous utilisons pour construire les modèles. Les algorithmes de Data Mining sont utilisés pour construire ces modèles. Il s'agit d'algorithmes conçus à partir de données qu'on subdivise en deux sous-ensembles, l'ensemble d'apprentissage et de test. D'une part l'ensemble d'apprentissage est constitué de 85% des données et nous permet de construire l'algorithme, d'autre part, l'ensemble de test contient 15% des données et sert à mesurer la performance et la précision du modèle élaboré. Cet ensemble nous permet également de minimiser le sur-apprentissage et de régler les paramètres des algorithmes.

On parle de sur-apprentissage lorsque l'algorithme a enclin d'apprendre profondément les données qui lui permettent l'apprentissage. On dit qu'il existe un problème de généralisation. En d'autres termes, l'algorithme ne vas pas aboutir à des prédictions rigoureuses en ce qui concerne les résultats des nouvelles données encore non étiquetées. Durant la modélisation et pour chaque algorithme testé, on a recours à la classification. Ceci nous informe sur la nature de la variable cible, étant une variable de nature catégorielle.

1 Modèles binaires

Dans cette section, on s'intéressera à la modélisation de la variable **Rating** en prenant en considération deux modalités :

- Catégorie d'investissement.
- Catégorie spéculative.

1.1 Pré-traitement de la variable cible

La variable **Rating** dans notre base de données a comme modalités :

```
[3]  data.Rating.value_counts()

     B      181
     BBB    147
     BB     125
     AA     111
     A       87
     AAA     59
     CCC     19
     Name: Rating, dtype: int64
```

FIGURE 5.1 – Fréquence des modalités dans notre variable cible

Pour qu'on puisse établir notre modélisation, il faut en premier lieu réorganiser les modalités en deux catégories.

```
[5]  data.Rating.value_counts()

     Inv    404
     Spe    325
     Name: Rating, dtype: int64
```

FIGURE 5.2 – Fréquence des modalités dans notre variable cible après modification

Ensuite, il faut attribuer à chaque modalité un code pour faciliter le travail des algorithmes de Machine Learning.

```
[9]  dict(zip(le.inverse_transform([0,1]),[0,1]))

     {'Inv': 0, 'Spe': 1}
```

FIGURE 5.3 – Dictionnaire des modalités

1.2 Sélection des variables

Dans le dernier chapitre, on a remarqué qu'on a de fortes corrélations entre les variables explicatives. Pour remédier à ce problème, on a choisi d'utiliser une fonction prédéfinie

de la bibliothèque Sklearn : **RFE**.

L'élimination récursive de caractéristiques (RFE) est une méthode de sélection de caractéristiques qui ajuste un modèle et élimine la ou les caractéristiques les plus faibles jusqu'à ce que le nombre spécifié de caractéristiques soit atteint. Les caractéristiques sont classées par les attributs coef_ ou feature_importances_ attributes, et en éliminant récursivement un petit nombre de caractéristiques par boucle, RFE tente d'éliminer les dépendances et la colinéarité qui peuvent exister dans le modèle.

RFE requiert un nombre spécifique de caractéristiques à conserver, mais il est souvent impossible de savoir à l'avance combien de caractéristiques sont valides. Pour trouver le nombre optimal de caractéristiques, la validation croisée est utilisée avec RFE pour évaluer différents sous-ensembles de caractéristiques et sélectionner la collection de caractéristiques ayant le meilleur score.

Grâce à cette fonction on a pu éliminer plusieurs variables. Ainsi, on a réussi à résoudre deux problèmes majeurs : la corrélation et le sur-ajustement.

1.3 Résultats trouvés

1.3.1 Régression Logistique

	precision	recall	f1-score	support
0	0.93	0.95	0.94	346
1	0.94	0.92	0.93	273
accuracy			0.94	619
macro avg	0.94	0.93	0.93	619
weighted avg	0.94	0.94	0.94	619

FIGURE 5.4 – Le résultat du modèle : Régression logistique

1.3.2 Support Vector Machine

	precision	recall	f1-score	support
0	0.95	0.96	0.96	346
1	0.95	0.94	0.94	273
accuracy			0.95	619
macro avg	0.95	0.95	0.95	619
weighted avg	0.95	0.95	0.95	619

FIGURE 5.5 – Le résultat du modèle : Régression logistique

Nous remarquons que la méthode du SVM est meilleure que celle de la régression logistique en termes de précision et de scoring.

2 Evaluation

2.1 La courbe ROC

Pour mesurer les performances d'un modèle de classification, la courbe ROC est une solution optimale pour avoir un résultat crédible à travers les seuils de classification. Avant de parler sur cette courbe, on doit bien définir les notions suivantes :

Quand le modèle prédit correctement la classe positive, on parle d'un résultat « **vrai positif** ». De façon analogue, quand le modèle prédit correctement la classe négative, on parle donc d'un résultat « **vrai négatif** » Or, un **faux positif** est un résultat où le modèle prédit incorrectement la classe positive. Et un **faux négatif** est un résultat où le modèle prédit incorrectement la classe négative.
La courbe ROC se base sur :

- Taux de vrais positifs : $TVP = \frac{VP}{VP + FN}$
- Taux de faux positifs : $TFP = \frac{FP}{FP + VN}$

Pour calculer les points d'une courbe ROC, il est plus efficace de calculer l'aire sous cette courbe, ou AUC, grâce à un algorithme de tri.

2.2 La courbe AUC

AUC signifie "aire sous la courbe ROC". Cette valeur mesure l'intégralité de l'aire à deux dimensions située sous l'ensemble de la courbe ROC.

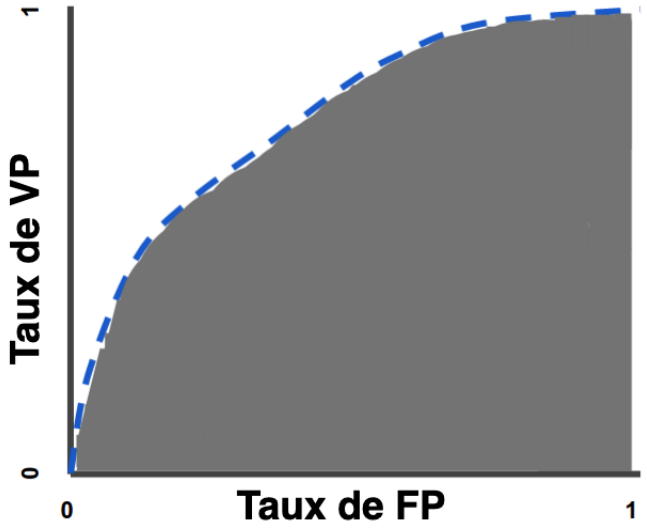

FIGURE 5.6 – Les courbes : ROC et AUC

Les valeurs d'AUC appartiennent à l'intervalle [0,1]. On a deux cas possibles :

- Si les prévisions sont erronées à 100%, le modèle a un AUC de (0,0)
- Si les prévisions sont correctes à 100%, le modèle a un AUC de (1,0)

Par exemple :

FIGURE 5.7 – Résultat du modèle de regression logistique

Parmi les avantages de L'AUC :

- L'AUC est invariante d'échelle : elle mesure la qualité du classement des prédictions, plutôt que leurs valeurs absolues.
- L'AUC est indépendante des seuils de classification : elle mesure la qualité des précisions du modèle quel que soit le seuil de classification sélectionné.

2.3 Application

On voit clairement la puissance du modèle "SVM", d'un coefficient de $\boxed{AUC = 0.95 = 95\%}$

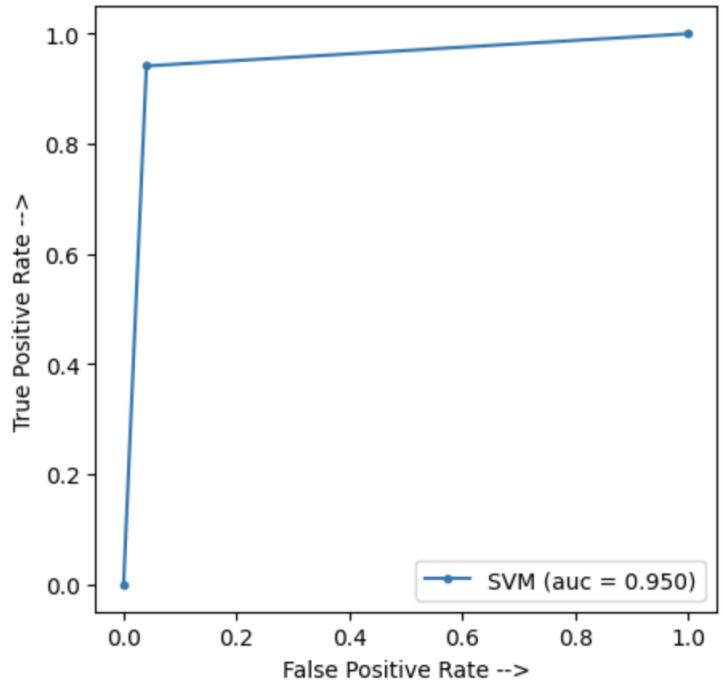

FIGURE 5.8 – Courbe AUC du modèle : SVM

Le modèle de la régression loqistique a : $\boxed{AUC = 0.933 = 93.3\%}$

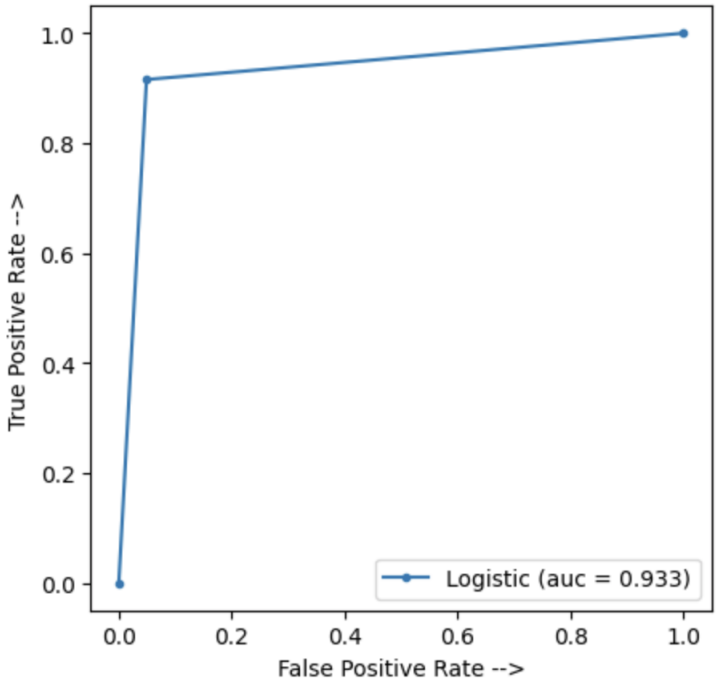

FIGURE 5.9 – Courbe AUC du modèle : Régression Logistique

2.4 Comparaison entre les modèles

Nom	Type de problème	Risque de sur-ajustement	Précision prédictive	Vitesse d'apprentissage	Fonctionne avec un grand nombre d'observations	Manipulation des variables non-pertinentes
RL	Classification	Fort	Faible	Rapide	Oui	Non
SVM	Classification & régression	Faible	Forte	Moyenne	Non	Oui

TABLE 5.1 – Etude comparative des deux algorithmes de prédiction

3 Modèles multiclass

Dans cette section, on s'intéressera à la modélisation de la variable **Rating** en prenant en considération toutes ses modalités :

- AAA
- AA
- A
- BBB
- BB
- B
- CCC

3.1 Préparation de la base de données

On a attribué à chaque modalité un code comme on a déjà fait dans la première partie.

```
dict(zip(le.inverse_transform([0,1,2,3,4,5,6]),[0,1,2,3,4,5,6]))
```

```
{'A': 0, 'AA': 1, 'AAA': 2, 'B': 3, 'BB': 4, 'BBB': 5, 'CCC': 6}
```

FIGURE 5.10 – Dictionnaire des modalités

Ensuite pour la sélection des variables, on utilise la même fonction utilisée dans le cas binaire. Cependant, cette fois-ci le modèle qui sera ajusté va être entraîner sur une variable cible ayant plus que 2 modalités.

Après cette étape, nos variables explicatives se réduisent à 39 variables.

3.2 Résultats trouvés

En se basant sur **"accuracy"**, on trouve que pour la régression logistique on a un score de :

$$\boxed{Train\_acc = 0.809 = 80.9\% \text{ et } Test\_acc = 0.7 = 70\%}$$

Tandis que pour le modèle SVM, on trouve :

$$\boxed{Train\_acc = 0.859 = 85.9\% \text{ et } Test\_acc = 0.75 = 75\%}$$

On ne peut pas se baser sur ces résultats, vu que les modalités ne sont pas réparties identiquement.On utilise **ROC_AUC** pour trouver des résultats pertinents.

3.2.1 Matrice de confusion RL

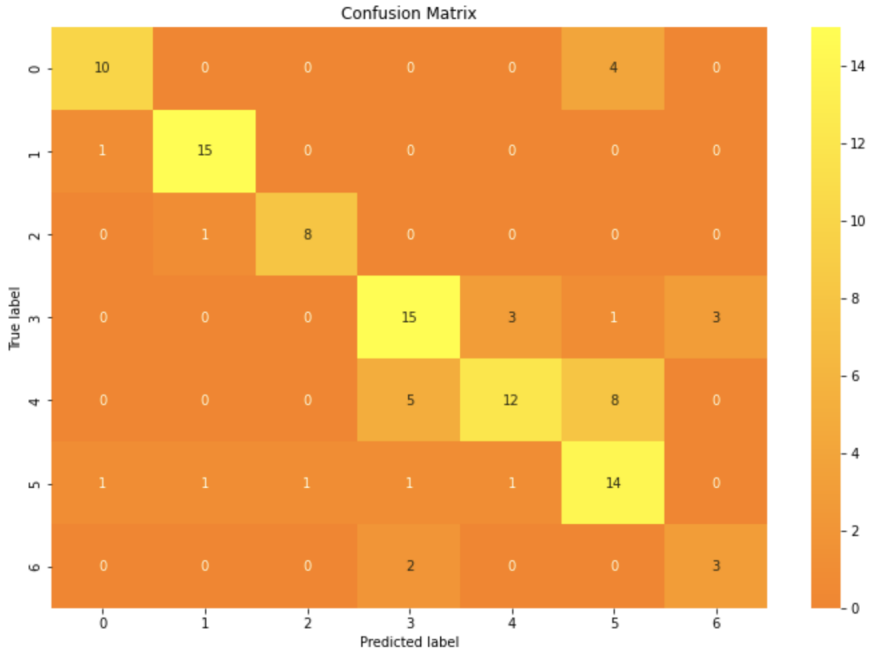

FIGURE 5.11 – Matrice de confusion du modèle régression logistique

3.2.2 Matrice de confusion SVM

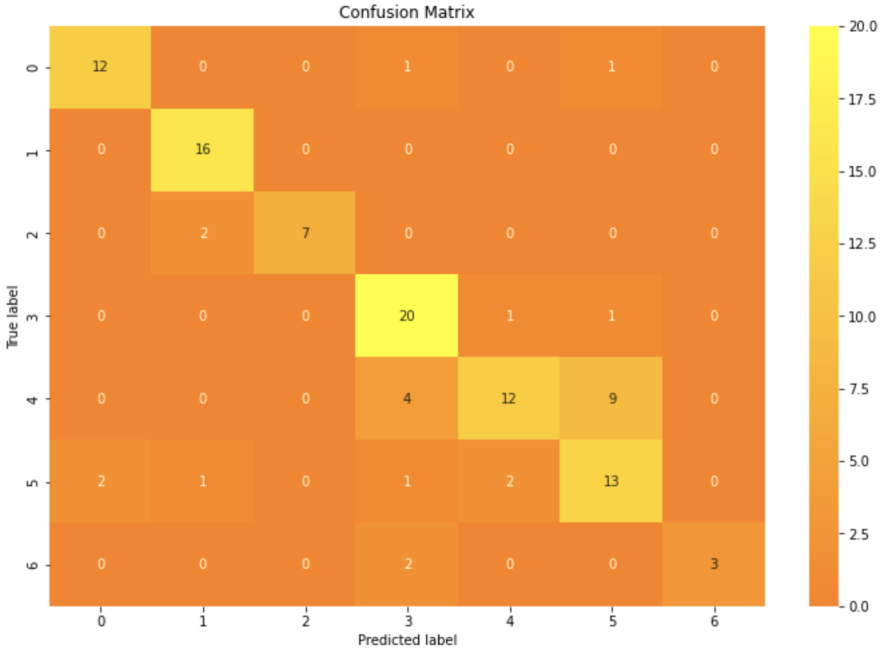

FIGURE 5.12 – Matrice de confusion du modèle SVM

En se basant sur **"ROC_AUC"**, on trouve que pour la régression logistique on a un score de :

$$\boxed{roc\_auc = 0.929 = 92.9\%}$$

Tandis que pour le modèle SVM, on trouve :

$$\boxed{roc\_auc = 0.963 = 96.3\%}$$

Conclusion

La désintermédiation financière a rendu la mission des agences incontournable. L'information émise par les agences est nécessaire pour la fixation du taux d'intérêt sur le marché des obligations souveraines. Mais elle n'est pas l'unique variable explicative.

Aujourd'hui, la dichotomie entre pays sûrs et pays risqués au sein de l'univers d'investissement est un élément plus décisif que la notation elle-même. De plus, la notation des agences doit s'accompagner d'une contre-expertise interne menée par les équipes économiques. En confrontant les points de vue, les acheteurs de dettes souveraines disposeront d'une information complète pour satisfaire leurs décisions d'investissement.

Dans le cadre de ce travail, nous avons étudié les déterminants de la notation financière souveraine attribuée par SP. Ce travail se réfère au cadre pratique considérant que les notes financières souveraines accordées par les agences de notation sont une mesure de la solvabilité internationale d'un État et que, de ce fait, ces agences mènent une analyse multidimensionnelle afin d'apprécier la capacité de cet État à servir sa dette extérieure vis-à-vis de ses créditeurs commerciaux.

Bibliographie

1. What is the Purpose of Data Science ? Know Its Importance. DataFlair
   https ://data-flair.training/blogs/purpose-of-data-science/
2. What Is Machine Learning ? NetApp
   https ://www.netapp.com/artificial-intelligence/what-is-machine-learning/
3. When Should I Use Regression Analysis ? Statistcs By Jim
   https ://statisticsbyjim.com/regression/when-use-regression-analysis/
4. Classification - Machine Learning. SimleLearn
   https ://www.simplilearn.com/classification-machine-learning-tutorial
5. Classification : Basic Concepts, Decision Trees, and Model. Kumar
6. Logistic Regression ó Detailed Overview. Saishruthi Swaminathan
   https ://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
7. DECODING THE POPULARITY OF JUPYTER AMONG DATA SCIENTISTS
   https ://www.analyticsinsight.net/decoding-popularity-jupyter-among-data-scientists/
8. Why Python Programming Language is important in Data Science ? Aaksh Kumar
   https ://medium.com/javarevisited/why-python-programming-language-is-important-in-data-science-beb4a7f91f75
9. Evaluating Multi-Class Classifiers
   https ://medium.com/apprentice-journal/evaluating-multi-class-classifiers-12b2946e755b
10. Deep Dive Into Multiclass Classification
    https ://www.kaggle.com/shrutimechlearn/deep-dive-into-multiclass-classification

Annexe
Le fonctionnement du SVM

Nous prenons le cas le plus aisé à représenter : la séparation en deux populations de m observations (classification binaire, donc) selon deux dimensions. On remarque qu'il existe une infinité de solutions possibles. Les trois modèles 'a', 'b' et 'c' offrent exactement les mêmes scores de classification sur cet échantillon. Ce qui les différencie, c'est leur capacité à généraliser. En effet, si le nombre d'observations augmente, il est fort probable que le modèle 'a' commence à produire des faux négatifs et 'c' des faux positifs. Le modèle 'b' semble plus optimal, justement parce qu'il maximise ses chances de généraliser. Cette représentation donne une intuition de la notion de « marge ». C'est l'une des idées essentielles du Support Vector Machine (SVM).

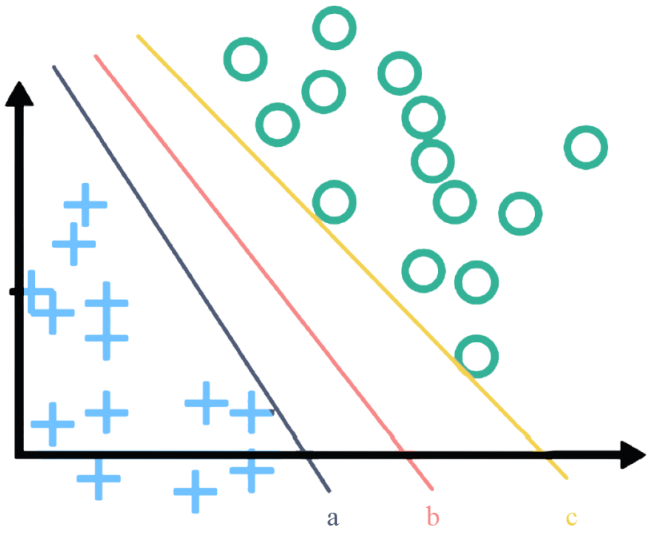

FIGURE 5.13 – Tous Modèles de classification possible

Soient :

- d_+ : la marge pour les exemples positifs comme étant la plus petite distance entre l'ensemble des exemples positifs et l'hyperplan

- d_- : la marge pour les exemples négatifs comme étant la plus petite distance entre l'ensemble des exemples négatifs et l'hyperplan.
- Θ : le vecteur des paramètres de notre modèle.
- La marge pour l'ensemble des données d'apprentissage est donc : $S = d_+ + d_-$

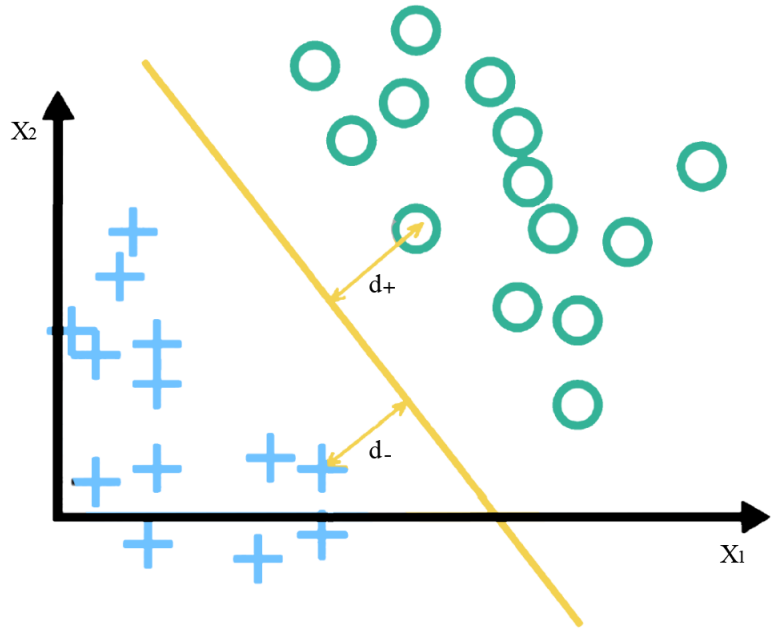

FIGURE 5.14 – La marge

Nous rappelons qu'en classification binaire, c'est le signe de $\Theta^T x_i$ qui donne la prédiction pour l'observation x_i

Imaginons que nous cherchions à définir un espace dans lequel nos données d'entraînement seraient toutes linéairement séparables. Il existe un hyperplan P qui sépare les observations positives et négatives tel que :

$$\forall x_i, x_j \in P, \Theta^T (x_i - x_j) = 0$$

D'un point de vue géométrique, cela est équivalent à dire que Θ et le vecteur défini par $x_i - x_j$ sont orthogonaux, comme nous nous en rendons compte dans la figure [5.33]. Nous y avons ajouté l'observation x_n , définie comme le point le plus proche de l'hyperplan (et donc celui qui servira à définir la marge).

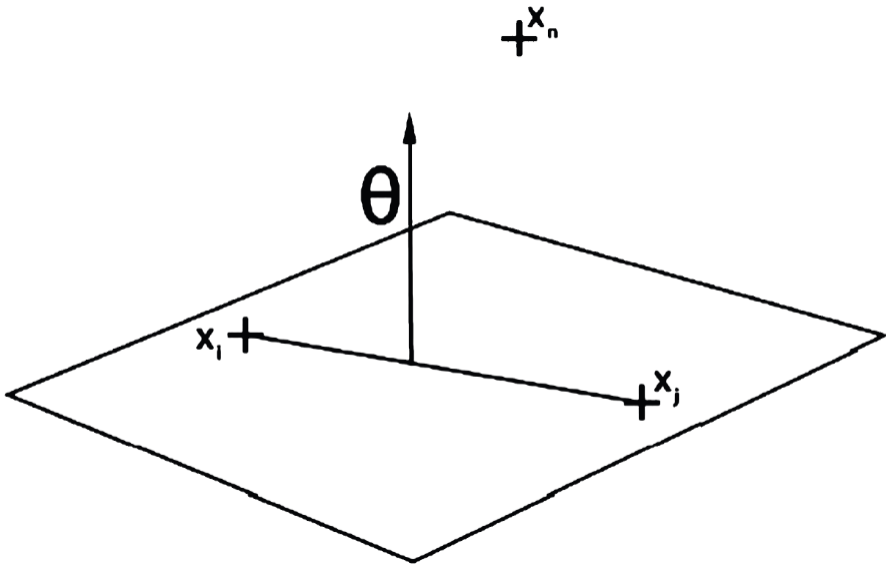

FIGURE 5.15 – Le point le plus proche de l'hyperplan

La distance entre x_n et l'hyperplan est définie par :

$$d = \frac{1}{\|\Theta\|} \left|\Theta^T (x_n - x)\right| \forall x \in P$$

Pour l'établir, il suffit d'écrire le produit scalaire entre Θ et $x_n - x$ et de décomposer ce dernier vecteur en sa composante du plan et sa composante orthogonale au plan, qui est précisément d.

L'hyperplan qui est parallèle à P et qui passe par x_n vérifie naturellement :

$$\Theta^T x_i \geq d \forall x_i \in S_+$$

Notons enfin que nous pouvons, sans rien changer à notre problème, normaliser le vecteur Θ de la sorte : $|\Theta^T x_n| = 1$ pour x_n , le point le plus proche de l'hyperplan séparateur. Ce sera une contrainte que nous imposerons au modèle. Dans ces conditions, la distance d, qu'il faut maximiser, devient simplement :

$$d = \frac{1}{\|\Theta\|}$$

Le problème que doit résoudre le SVM est donc celui-ci :

$$\begin{cases} \max\left(\frac{1}{\|\Theta\|}\right) \\ \text{SC}: \ \min_{i=1,2,\ldots,m} |\Theta^T x_i| = 1 \end{cases}$$