Elizabeth Wolfsohn, Amele Olivia Adamah, Etienne De Poix

Projet Applicatif de Sciences de données et Apprentissage: Prédiction de Survie d'une start-up

Résumé

Cette étude porte sur la prédiction de survie des start-ups en se basant sur ses données administratives ainsi que sur ses bilans financiers annuels. Plusieurs méthodes sont mises en œuvre, notamment la régression logistique avec sélection de variables, les arbres de décision, les Random Forests, ainsi que sur les réseaux de neurones. Les performances de ces méthodes seront évaluées et comparées entre elles afin d'identifier les modèles les plus pertinents dans le cadre de cette étude.

Introduction

De nos jours, le concept de start-up gagne en popularité dans le monde du travail pour son caractère novateur et dynamique. Ainsi, en raison de son caractère expérimental et de la recherche constante d'innovation, une start-up se distingue d'une entreprise traditionnelle, qui cherche principalement à optimiser son business model. [1] C'est pourquoi son comportement diffère au cours du temps de celui d'une entreprise classique. Dès lors, nous cherchons à déterminer comment prédire la survie d'une start-up à partir des données financières de ses premières années d'existence et à partir de ses caractéristiques administratives.

Une start-up est principalement définie comme étant une entreprise "innovante de création récente". [2] Etant donné que cette notion peut être très large, notamment en termes de secteurs car l'innovation peut techniquement avoir lieu dans n'importe quel secteur, ainsi qu'en termes d'âge de l'entreprise, nous nous sommes essentiellement concentrés sur l'étude des premières années des entreprises spécialisées dans les secteurs de l'informatique et de la R&D. Etant donné que nous avons relativement peu d'informations précises sur les entreprises dans un temps proche de la création de l'entreprise, certains jeux de données comportent des informations récentes, comme les données administratives. Mais concernant les données financières, nous nous sommes concentrés sur les trois premières années de vie de l'entreprise, dont l'année de sa création.

Afin de prédire la survie d'une start-up, nous avons utilisé trois jeux de données, tous extraits du site data.gouv : le premier produit par l'Insee et centré sur les données administratives des entreprises [3], le deuxième produit également par l'Insee et centré sur les données administratives des établissements des entreprises [3], et enfin le dernier produit par Signaux Faibles, qui réutilise les données de bilan de l'INPI, et centré sur les données financières des entreprises [4]

Nous analyserons d'abord les différents jeux de données exploités ainsi que la mise en forme et la préparation du jeu de données final, réalisées par Elizabeth Wolfsohn. Ensuite, nous ferons l'analyse des relations entre les données du jeu de données final, menée par Etienne De Poix. Enfin, nous étudierons l'implémentation des divers modèles d'apprentissage et nous discuterons des résultats produits, ainsi que de la comparaison des performances entre les modèles. Cette étude porte à la fois sur des modèles hors réseaux de neurones, dont la régression logistique pénalisée (Lasso), les arbres

de décision et les Random Forests, implémentés par Amele Olivia Adamah, ainsi que des modèles de réseaux de neurones à différentes densités, implémentés par Etienne De Poix.

1 Mise en forme et préparation des jeux de données

1.1 Présentation des jeux de données initiaux et finaux

Composé de 29 000 000 instances environ, le jeu de données *Stock Unité Légale* contient les données administratives de toutes les entreprises de France. C'est ce jeu de données qui comporte la sortie à prédire (***etatAdministratifUniteLegale***), indiquant si une entreprise est active ou a cessé d'exister. Il contient également des données essentielles concernant l'extraction des start-ups. En particulier, les variables ***activitePrincipaleUniteLegale*** et ***categorieJuridiqueUniteLegale*** nous permettent de sélectionner les entreprises qui ont des caractéristiques typiques de start-ups, dont respectivement la catégorie juridique SAS (Sociétés par actions simplifiées) ainsi que les secteurs considérés comme étant "innovants", comme par exemple la R&D et l'informatique. [5] Ce jeu de données contient une autre information essentielle qu'est la date de création de la start-up ***(dateCreationUniteLegale)*** car elle nous servira à sélectionner les bilans financiers des 3 premières années de chaque start-up dans le jeu de données *Ratios INPI BCE*. Enfin, il contient d'autres données susceptibles d'être intéressantes comme la ***trancheEffectifsUniteLegale*** qui contient le nombre de personnes dans la start-up ou encore ***nombrePeriodesUniteLegale*** qui contient le nombre de fois que la start-up a changé de statut au niveau d'un certain aspect. Il contient également des valeurs non explicatives, comme des prénoms ou bien des nomenclatures, que nous avons choisies de supprimer.

Composé de 60 000 000 instances environ, le jeu de données *Stock Unité Établissement* contient toutes les informations administratives relatives aux établissements (les filiales et la maison mère) associées aux entreprises. Il contient diverses informations comme les données de géolocalisation. Il comporte également la variable ***etatAdministratifUniteLegale*** pour chaque entreprise. Ainsi, nous avons trouvé intéressant d'enrichir notre jeu de données en comptant le nombre d'établissements associé à chaque start-up. Afin de ne pas rendre l'étude trop complexe, nous avons choisi de ne garder que la variable qui donne le nombre de d'établissements de chaque start-up afin de préserver l'étude à l'échelle des start-ups et non de ses filiales.

Le jeu de données *Ratios INPI BCE* contient les données financières de diverses entreprises pour chaque année déclarée. Il comporte 23 variables dont 20 variables financières, comme le chiffre d'affaires ou bien le résultat de l'exercice (bénéfice ou perte). Les trois autres sont la date à laquelle le bilan financier a été déclaré, le siren ainsi que le type de bilan (s'il a été simplifié ou complet). A l'aide de ces variables, nous cherchons à extraire les bilans financiers sur les 3 premières années de vie de chaque start-up présente et de calculer pour chaque variable financière la moyenne, l'écart-type ainsi que le taux de croissance annuel moyen (TCAM). Cela nous permettra de mieux mettre en valeur les relations entre les 3 valeurs basées sur chaque variable et de réduire le nombre de valeurs manquantes.

Le jeu de données final est la synthèse de trois jeux de données différents, qui sont tous reliés par l'identifiant unique des entreprises (siren). Il comporte 36 388 instances avec 68 variables, contenant les données administratives et enrichi de nouvelles variables financières calculées à partir des valeurs brutes ainsi que le nombre d'établissements de chaque start-up.

1.2 Processus de traitement des jeux de données

1.2.1 Jeu de données administratives (Stock Unité Légale)

Nous commençons le processus de préparation de données par le traitement du bilan administratif des entreprises. En effet, étant donné qu'il contient les données essentielles telles que la variable cible ***etatAdministratifUniteLegale*** (l'éventuelle survie d'une entreprise), le siren associé, ainsi que les différentes caractéristiques spécifiques à une start-up, comme le secteur ou la catégorie juridique, ce jeu de données est déterminant en ce qui concerne la pré-sélection des start-up étudiées. Par conséquent, nous supprimons toutes les lignes contenant des valeurs manquantes au niveau de la variable ***etatAdministratifUniteLegale***. Nous faisons de même pour le siren afin de pouvoir relier les données de l'entreprise provenant des différents jeux de données. Nous supprimons également les lignes dont la date de création est inconnue, car elle est nécessaire afin de récupérer les bilans financiers des premières années de vie de la start-up. De plus, nous supprimons toutes les variables non explicatives qui servent seulement à dénommer l'entreprise, comme les variables ***prenom1UniteLegale*** ou bien ***denominationUsuelle1UniteLegale***. Les variables ne comportant que des variables manquantes sont supprimées, comme par exemple ***caractereEmployeuUniteLegal****e*.

Nous filtrons ensuite les entreprises de telle sorte à ce qu'elles respectent les critères pouvant caractériser une start-up. En particulier, concernant les catégories juridiques (***categorieJuridiqueUniteLegale***) , nous n'avons retenu que les SAS (Sociétés par actions simplifiées), identifiés par le numéro 5710. [5] Par ailleurs, concernant les secteurs (***activitePrincipaleUniteLegale***), nous avons retenu une sélection de secteurs centrés sur les secteurs de l'informatique, de la technologie et du R&D. [6]

Les variables catégorielles sont ensuite encodées de sorte à ce qu'elles puissent être traitées par les modèles. Initialement les valeurs associées à la variable ***activitePrincipaleUniteLegale*** sont caractérisées de la manière suivante par exemple : "70.22Z". Nous avons encodé cette variable en extrayant les 4 chiffres contenus dans la valeur initiale. Les valeurs de la variable ***categorieJuridiqueUniteLegale*** et ***trancheEffectifsUniteLegale*** sont simplement transformées en entier. Les autres variables catégorielles ne sont composées que de 2 ou 3 valeurs distinctes, comme les variables ***etatAdministratifUniteLegale*** ou bien ***categorieEntreprise***. Nous avons dans ce cas procédé à un mapping à l'aide d'un dictionnaire composé manuellement et nous avons comblé les éventuelles valeurs manquantes par la valeur 0.

1.2.2 Jeu de données administratives des établissements (Stock Unité Établissement)

Nous filtrons les sirens de sorte à ce que ce jeu de données ne contient que les sirens du jeu de données administratives. Afin de garder une perspective centrée sur les start-ups en général et non sur les établissements en particulier, nous créons une nouvelle variable et nous retiendrons que celle-ci de ce jeu de données. Cette variable est basée sur le nombre d'établissements pour chaque start-up. Si une start-up a fermé, le nombre d'établissements associés reste quand même non nul. Par conséquent, cette variable ne révèle pas la survie ou non d'une start-up.

1.2.3 Jeu de données sur les bilans financiers annuels des entreprises (Ratios Inpi BCE)

De la même manière que pour le jeu de données sur les établissements associés aux entreprises, nous avons filtré les sirens de sorte à ce que le jeu de données financières ne contient que les sirens contenus dans le jeu de données administratives des entreprises. Après filtrage, nous avons alors découvert que seulement 12% des 280 000 entreprises (donc 36 000 entreprises environ) ont déclaré au moins un bilan financier dans ce jeu de données. Dans le jeu de données administratives filtré, le taux de survie des start-ups est de 74%, tandis que parmi les start-ups présentes dans le jeu de données administratives et ayant déclaré au moins un bilan financier de n'importe quelle année, ce taux de survie atteint 82%. Or, les données financières peuvent être très pertinentes dans l'évaluation de la survie d'une start-up. C'est pourquoi nous avons choisi de nous limiter aux start-ups ayant déclaré au moins un bilan financier dans un certain intervalle d'années situé au début de la vie de l'entreprise. Afin d'avoir un taux de survie le plus similaire possible à celui des start-ups du jeu de données administratives filtré, nous avons choisi l'intervalle des 3 premières années de l'entreprise (dont l'année de création de celle-ci). Avec ces critères, nous obtenons ainsi un taux de survie de 76% (contre 74% dans le jeu de données administratives filtré qui incluent les start-ups qui ont déclaré ou non un bilan financier). Dans cette version, étant donné que chaque instance est déterminée par le siren et l'année du bilan financier associé, nous avons retiré toutes les instances qui concernent les bilans financiers postérieurs à cet intervalle. Pour chaque variable financière, nous avons calculé la moyenne des 3 valeurs, son taux de croissance annuel moyen (TCAM) ainsi que son écart-type. [7]

La stratégie pour les calculer selon le nombre de valeurs existantes est la suivante :

- si au moins une valeur existe sur les 3, seule la moyenne peut être calculée
- si au moins deux valeurs existent, nous pouvons calculer également l'écart-type et le TCAM
- si les trois valeurs sont manquantes, les trois variables ne peuvent pas être calculées pour cette start-up

L'avantage du TCAM est qu'il est adaptable en termes au nombre de périodes renseignées et n'a besoin que de deux valeurs (la valeur initiale et la valeur finale). Bien qu'il contienne des informations importantes, le TCAM possède certaines contraintes : il n'accepte pas les valeurs initiales nulles (car division par 0) ainsi que les valeurs de signe différents. Néanmoins, la moyenne et l'écart-type permettent d'apporter des informations complémentaires au TCAM.

Après le processus de calculs pour tout le jeu de données, la fusion des trois jeux de données et la constitution des train set, validation set et test set, nous comblons les valeurs financières manquantes en y plaçant la médiane de chaque variable associée à la moyenne et à l'écart-type respectivement provenant du train set, afin d'éviter la fuite de données. Concernant les TCAM, nous comblons les valeurs manquantes par 0, pour signifier qu'il n'y a pas de croissance pour une certaine start-up.

Afin de ne pas avoir d'entreprises trop anciennes tout en ne supprimant pas d'instances qui ont déclaré un bilan financier dans les 3 premières années, nous n'avons gardé que les start-ups dont l'année de

création est supérieure ou égale à 2009. Ensuite, nous avons retiré la variable de date de création car nous ne voulons pas que le modèle ne puisse prédire que la survie des start-ups créées entre 2009 et 2025. Ainsi, le taux de survie de ces start-ups descend légèrement à 75%, ce qui se rapproche plus du taux de survie de toutes les start-ups confondues (74%). De plus, nous retirons toutes les variables dont la variance est nulle, ce qui est le cas seulement pour ***categorieJuridiqueUniteLegale***. Enfin, nous retirons la variable siren car il ne s'agit que d'une variable d'identification. Après la séparation de l'ensemble d'entraînement, l'ensemble de validation et celui de test et après avoir comblé les valeurs manquantes, nous procédons à l'application des différents modèles d'apprentissage automatique que nous avons sélectionnés.

2. Présentation des méthodes de prédiction

L'objectif est de prédire la survie d'une start-up à partir de ses caractéristiques financières et administratives. La variable cible est binaire : une start-up est soit **active (1)**, soit **non active (0)**. Il s'agit donc d'un problème de **classification binaire**.

Afin d'aborder cette problématique, plusieurs modèles de prédiction ont été comparés. Quatre approches ont été retenues : **la régression logistique**, **l'arbre de décision**, **le Random Forest et le réseau dense**. Ces modèles ont été choisis afin d'analyser les compromis entre **simplicité**, **interprétabilité** et **capacité de généralisation**, et d'identifier la méthode la plus adaptée à la prédiction de la survie des start-ups.

2.1 Régression logistique

2.1.1 Justification du choix

La régression logistique a été choisie comme méthode de base pour aborder la problématique de prédiction de la survie des start-ups. Deux arguments principaux justifient ce choix. Tout d'abord, la variable cible est binaire. La régression logistique est adaptée à ce type de prédiction en raison de la fonction de sigmoid qu'il est possible d'utiliser pour la classification, renvoyant directement une probabilité. Ensuite, cette façon de faire est très interprétable, et permet d'obtenir des relations claires entre les données.

2.1.2 Résultats et interprétation

Explication des résultats :

La régression logistique obtient une **accuracy de 0.8976**, indiquant qu'environ **89,8 % des start-ups sont correctement classées**. Le score **ROC AUC de 0.9576** témoigne d'une **excellente capacité de discrimination**, confirmant que le modèle distingue efficacement les start-ups survivantes des start-ups non actives.

L'analyse du rapport de classification montre que :

- La classe des start-ups survivantes (classe 1) est très bien prédite, avec une précision de 0.91, un rappel de 0.96 et un F1-score de 0.93, ce qui indique que le modèle identifie correctement la grande majorité des entreprises qui parviennent à survivre.
- La classe des **start-ups non actives (classe 0)** présente un **rappel plus faible (0.71)**, ce qui signifie qu'une partie des entreprises vouées à disparaître est encore classée comme survivante. Toutefois, la **précision de 0.85** montre que, lorsque le modèle prédit une non-survie, cette prédiction est généralement correcte.

Ces résultats indiquent que la régression logistique capture efficacement les tendances globales, mais reste limitée pour détecter les cas les plus difficiles.

Les scores obtenus sur les deux classes mettent en évidence un déséquilibre dans la qualité de prédiction entre les deux classes. Le modèle est particulièrement performant pour identifier les start-ups survivantes, avec un nombre élevé de vrais positifs et très peu de faux négatifs. En revanche, la détection des start-ups non actives reste plus difficile, comme l'indique le nombre relativement élevé de faux positifs.

2.2 Arbre de décision

2.2.1 Justification du choix

L'arbre de décision a été testé afin d'explorer une approche non linéaire et interprétable pour la prédiction de la survie des start-ups. Ce modèle permet de capturer des interactions plus complexes entre les variables financières et administratives. Les arbres de décision permettent de gérer des données hétérogènes (variables financières, catégorielles, administratives), ce qui en fait un outil robuste pour analyser des start-ups issues de secteurs variés.

2.2.2 Décisions prises et justification

Afin de limiter le surapprentissage, plusieurs paramètres de l'arbre de décision ont été contrôlés. La profondeur maximale de l'arbre a été fixée à 6, ce qui permet de restreindre la complexité du modèle tout en conservant une capacité suffisante à modéliser des relations non linéaires.
Le paramètre ***min_samples_leaf*** = 30 a été choisi afin d'imposer un nombre minimal d'observations dans chaque feuille, réduisant ainsi la sensibilité du modèle au bruit des données.
Ces choix visent à obtenir un arbre plus stable et capable de généraliser sur les données de test.

2.2.3 Résultats et interprétation

L'arbre de décision obtient une accuracy de 0.928, indiquant une très bonne performance globale sur l'ensemble des 10 917 start-ups analysées. Le score ROC AUC de 0.979 traduit une excellente capacité de discrimination, supérieure à celle observée avec la régression logistique.

L'analyse du rapport de classification montre que :

- La classe des start-ups survivantes (classe 1) est prédite avec une précision de 0.94, un rappel de 0.97 et un F1-score de 0.95, ce qui confirme une détection très efficace des entreprises encore actives.

- La classe des start-ups non actives (classe 0) présente également de bonnes performances, avec un rappel de 0.81, nettement supérieur à celui obtenu avec la régression logistique. Cela indique une meilleure capacité du modèle à identifier les entreprises vouées à l'échec.

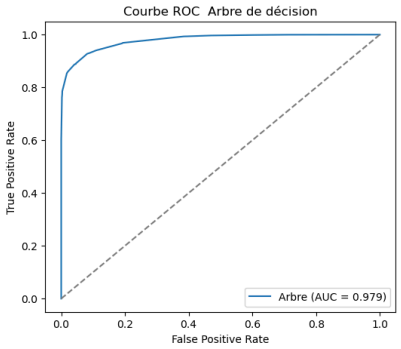

figure 1 : Courbe ROC

La courbe ROC associée à l'arbre de décision met en évidence une excellente capacité du modèle à distinguer les classes. La courbe est nettement située au-dessus de la diagonale, ce qui confirme la pertinence du modèle pour distinguer les start-ups survivantes des start-ups non actives.

Avec un score **AUC de 0.979**, l'arbre de décision démontre une performance très élevée, traduisant un taux de vrais positifs élevé pour de faibles taux de faux positifs. La forme de la courbe, proche du coin supérieur gauche, indique que le modèle parvient à identifier efficacement les start-ups à risque tout en limitant les erreurs de classification.

Ces résultats montrent que l'arbre de décision parvient à mieux capturer les structures non linéaires et les interactions entre variables, ce qui améliore significativement la détection des start-ups non actives tout en maintenant d'excellentes performances globales.

2.3 Random Forest

2.3.1 Justification du choix

Le Random Forest a été utilisé afin de dépasser les limites des arbres de décision simples, notamment leur forte variance et leur sensibilité au surapprentissage. Cette méthode d'ensemble repose sur l'agrégation de plusieurs arbres de décision, ce qui permet de capturer des relations non linéaires complexes tout en améliorant la robustesse et la capacité de généralisation du modèle.

Dans le cadre de la prédiction de la survie des start-ups, le Random Forest est particulièrement pertinent car il peut exploiter simultanément un grand nombre de variables financières et administratives, tout en limitant l'impact du bruit et des fluctuations présentes dans les données. Il constitue ainsi un bon compromis entre performance prédictive et stabilité.

2.3.2 Décisions prises et justification

Plusieurs paramètres ont été définis afin de contrôler la complexité du modèle et d'éviter le surapprentissage. Le nombre d'arbres a été fixé à 300, ce qui permet d'obtenir des prédictions stables

grâce à l'effet de moyennage tout en conservant un temps de calcul raisonnable. La profondeur maximale des arbres a été limitée à 10, afin d'éviter des arbres trop complexes, tandis que le paramètre ***min_samples_leaf*** = 20 impose un nombre minimal d'observations par feuille, réduisant la sensibilité du modèle aux valeurs atypiques.

2.3.3 Résultats et interprétation

Le modèle Random Forest atteint une accuracy de 0.900, indiquant une bonne performance globale sur les données de test.
Le score ROC AUC de 0.97 met en évidence une excellente capacité de discrimination, confirmant que le modèle distingue efficacement les start-ups survivantes des start-ups non actives, indépendamment du seuil de décision.

L'analyse du rapport de classification montre que :

- La classe des start-ups survivantes est prédite, avec un **rappel** de **0.99** et un F1-score de **0.94**, ces valeurs indiquent que le modèle identifie presque toutes les start-ups survivantes, avec très peu de faux négatifs, ce qui est crucial pour ne pas passer à côté d'entreprises encore actives.
- La classe des start-ups non actives présente une **précision** élevée **(0.95)** mais un **rappel** plus faible **(0.63)**, le modèle est très précis lorsqu'il prédit qu'une start-up va disparaître, mais le rappel plus faible montre qu'un nombre significatif de start-ups vouées à l'échec n'est pas détecté.

Ces résultats traduisent un modèle performant mais orienté vers la minimisation des faux positifs pour la classe non actives, ce qui peut être pertinent dans un contexte de réduction du risque d'erreur de classification.

Importance des variables :

l'analyse de l'importance des variables révèle que les facteurs les plus déterminants sont liés à la **structure administrative** et **juridique** de l'entreprise. Les variables ***nombrePeriodesUniteLegale, categorieEntreprise et anneeCategorieEntreprise*** dominent largement en terme d'importance.

La taille de l'entreprise ***trancheEffectifsUniteLegale*** et le nombre d'établissements jouent un rôle secondaire, tandis que les indicateurs financiers(***Resultat_net_moyen***, ***Chiffre_d_affaires_moyen***) apparaissent comme peu influents dans ce modèle.

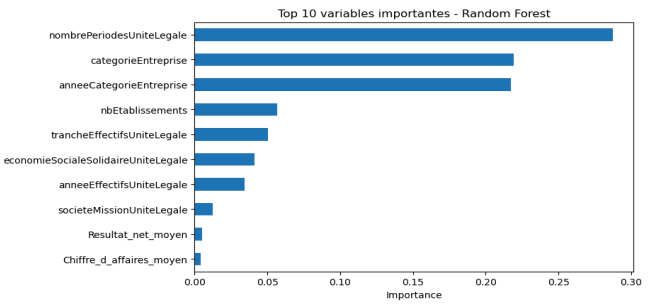

figure 2 : variables importantes pour la prédiction

2.4 Réseau de neurone dense

2.1.1 Justification du choix

Le recours au Deep Learning, et plus spécifiquement à un Perceptron Multicouche (MLP), a été motivé par l'ambition de capturer des interactions de haut niveau entre les variables, difficilement détectables par des méthodes classiques. Contrairement aux arbres de décision qui partitionnent l'espace de manière orthogonale, les réseaux de neurones opèrent des transformations non linéaires successives. Cette approche vise à extraire des représentations latentes complexes pour potentiellement surpasser la performance des méthodes d'ensemble sur les cas les plus ambigus.

2.3.2 Décisions prises et justification

Pour la création des modèles, le choix a été fait de se concentrer sur l'étude de deux paramètres clé, la taille des modèles et le dropout. Les paramètres de l'optimiseur et de la fonction de coût ont été déterminés selon les standards pour les réseaux denses, à savoir respectivement l'optimiseur Adam et la fonction de coût Binary Cross-Entropy (BCE). 4 modèles ont été testés, de manière à évaluer l'impact de leur taille et de leur dropout. Un modèle de base de 3 couches, un modèle "light" de 2 couches, un modèle "deep" de 4 couches, et un dernier modèle, identique au modèle "deep" mais comprenant en plus du dropout après chaque couche. Différents nombres d'époques ont été testés. Les résultats ont finalement été obtenus avec 20 époques.

2.3.3 Résultats et interprétation

Finalement, tous les modèles ont obtenu de bonnes performances. Les modèles base, light, deep et avec dropout ont obtenu respectivement un f1-score de 94.9%, 95.4%, 95.3% et 95.3%.
On remarque quelque chose d'étrange avec les résultats, en effet, ils sont très similaires malgré des modèles très différents. La courbe d'apprentissage permet de mieux comprendre. Elle est similaire pour tous les modèles.

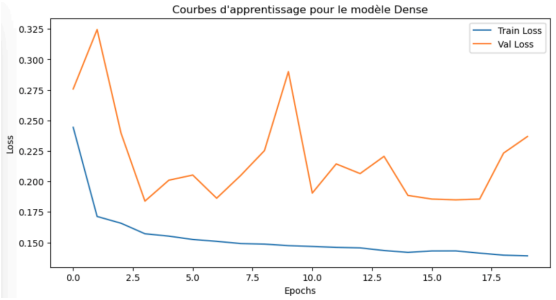

figure 3 : courbe d'apprentissage pour le réseau de neurone dense

On peut remarquer que le modèle n'apprend pas vraiment. Il finit son apprentissage au bout de 3 époques, ce qui est très peu, puis fait du surapprentissage. Finalement, le modèle apprend principalement un modèle linéaire, ce qui va dans le sens des résultats obtenus par les modèles d'arbres de décision et de régression logistique, qui ont obtenu des résultats très proches.

L'apprentissage extrêmement rapide (3 époques) et la similarité des scores suggèrent la présence de variables à fort pouvoir discriminant mais potentiellement liées à un biais de survie (ex: nombre de périodes administratives ou mise à jour des effectifs en 2025). Le modèle identifie ces 'proxys' de

survie quasi instantanément, rendant inutile l'extraction de caractéristiques financières plus profondes par les couches successives.

Cela peut aussi s'expliquer par la qualité médiocre des données financières. En effet, plus de deux tiers des entreprises n'ont déclaré qu'un seul bilan financier sur les 3 ans. Dès lors, il n'est pas possible de déterminer l'évolution de la start up dans le temps, un indicateur très fort de sa performance réelle.

3. Résultats

3.1 Comparaison des méthodes

Au final, les performances obtenues sur les modèles sont très élevées et très serrées pour les quatre modèles. La comparaison est faite sur le f1-score, qui représente leur performance globale.

Régression logistique : Ce modèle simple et interprétable détecte efficacement les start-ups survivantes avec un f1-score de 0.93, mais est légèrement moins performant pour identifier les start-ups fermées avec un f1-score de 0.77. Cela reflète la nature linéaire de la régression logistique et sa capacité à capturer les tendances globales.

L'Arbre de Décision (Simple) : Avec une accuracy de 92,8 % et un ROC AUC de 0,979, l'arbre de décision distingue bien les classes. La classe des start-ups survivantes est détectée avec un F1-score de 0,95, ce qui confirme sa capacité à capturer les relations non linéaires dans les données.

Random Forest : Ce modèle détecte presque toutes les start-ups survivantes avec un f1-score de 0.94, mais sous-détecte davantage les start-ups fermées.

Le réseau de neurone dense : Les réseaux avec le plus de couches affichent les meilleures performances, atteignant aussi un f1-score de 95.3%. Malgré un certain niveau de sur-apprentissage, le modèle obtient de très bonnes performances.

3.2 Analyse

Cette similarité dans les résultats montre que les relations entre les données ne présentent pas de relation cachée. Des méthodes statistiques simples, telles que l'arbre de décision ou la régression logistique suffisent à obtenir de très bons résultats.

Néanmoins, les modèles bénéficieraient d'une meilleure performance s'il y avait plus de données financières dans le jeu de données initial des bilans financiers. En effet, sur toutes les start-ups sélectionnées du jeu de données, moins de 20% d'entre elles ont déclaré un bilan financier sur les 3 premières années de leur existence. Ainsi, en ne gardant que les start-ups qui remplissent ces critères, il pourrait être argumenté qu'il y ait un biais sur leur survie. Nous avons alors pensé à une alternative consistant à garder toutes les startups du jeu de données administratives filtré (cela inclut même les entreprises qui n'ont déclaré aucun bilan financier), à ne pas étudier le contenu des bilans financiers et à ajouter plutôt dans le jeu de données final une variable qui indique si une certaine start-up a déclaré au moins un bilan financier dans les 3 premières années de vie de celle-ci (année de création incluse).

Mais, par conséquent, nous ne pourrions plus nous baser sur les données financières et nous n'aurions que des données récentes, qui sont les données administratives.

Conclusion

Ainsi, nous avons pu créer un jeu de données complet, permettant de prédire la survie d'une startup. Nous avons pu tester 4 modèles d'apprentissage simples (Régression logistique, arbre de décision) et plus complexes (Forêt aléatoire, réseau de neurone dense). Finalement, ces modèles obtiennent chacun de bons résultats sur le dataset, les meilleurs étant le réseau dense et l'arbre de décision avec un f1-score de 95%. Cependant, les bons résultats de ces modèles mettent aussi en évidence les relations simples entre les données pour la prédiction. Cela peut s'expliquer par la faible qualité des données financières, n'affichant souvent qu'une seule année d'activité. Pour aller plus loin, il serait intéressant de faire un modèle basé sur au moins 2 bilans financiers déclarés, de manière à pouvoir faire apprendre aux modèles l'impact de la croissance.

Bibliographie

[1] Qu'est-ce qu'une startup ? | Bpifrance Création

[2] Start-up — Wikipédia

[3] Base Sirene des entreprises et de leurs établissements (SIREN, SIRET)

liens de téléchargement :
https://www.data.gouv.fr/api/1/datasets/r/350182c9-148a-46e0-8389-76c2ec1374a3

https://www.data.gouv.fr/datasets/base-sirene-des-entreprises-et-de-leurs-etablissements-siren-siret?resource_id=a29c1297-1f92-4e2a-8f6b-8c902ce96c5f

[4] Jeu de données - Ratios Financiers (BCE / INPI) | data.gouv.fr

lien de téléchargement :
https://www.data.gouv.fr/api/1/datasets/r/9d213815-1649-4527-9eb4-427146ef2e5b

[5] Catégories juridiques | Insee

[6] Nomenclature d'activités française – NAF rév. 2 | Insee

[7] Calcul taux de croissance : formule + exemples Excel