Parcours Data Scientist OpenClassrooms

Rapport de stage en entreprise

https://openclassrooms.com/

Stage effectué par Morgan SCAO entre avril et juin 2018

Table des matières

1. Introduction

Lorsque l'on est une banque ou un assureur, lorsque l'on travaille en B to B, on peut parfois se demander si on sera effectivement payé à la fin d'un contrat, ou remboursé s'il s'agit d'un prêt…Eh oui il n'est pas toujours évident de faire confiance à une entreprise avec laquelle on pourrait travailler.

C'est dans ce contexte qu'arrivent les agences de notation. On en connait tous de célèbres qui notent même la France et tous les autres pays. On ne va pas aller jusque-là aujourd'hui, on se contentera de chercher à noter l'ensemble des entreprises françaises, ce qui est déjà pas mal.

Mais quelles informations a-t-on à notre disposition ?

Comment noter une entreprise qu'on ne connait pas ?

Et comment nos clients devront-ils interpréter cette note ?

Nous allons voir tout cela.

2. Contexte

Dans le cadre du projet 8 de ma formation Data Scientist dispensée par CentraleSupélec chez OpenClassrooms j'ai opté pour le stage en entreprise plutôt qu'une veille thématique. Le fait de se retrouver confronté avec une problématique concrète, des clients en face, des collègues autour, bref une vraie ambiance de travail me paraissait plus en adéquation avec ce que je recherchais dans le but d'être pleinement employable à la fin de ma formation.

L'entreprise dans laquelle je suis propose à ses clients des analyses complètes et approfondies à partir d'un numéro de SIREN. Le but étant de lever une alerte dès qu'une anomalie pouvant entrainer une radiation, et donc une insolvabilité future survient.

La surveillance se fait à partir de différentes sources de données publiques dont on peut avoir un aperçu global ici : https://www.data.gouv.fr/fr/

- INSEE
  - Base SIRENE : https://www.sirene.fr/sirene/public/accueil
- DILA (Direction pour l'Information Légale et Administrative)
  - BODACC (Bulletin officiel des annonces civiles et commerciales)
  - JOAFE (Journal Officiel des Associations et Fondations d'Entreprise)
  - BALO (Bulletin des annonces légales et obligatoires)
- INPI (Institut national de la propriété industrielle)
  - RNCS (Registre national du commerce et des sociétés)
  - Marques (Consultation en ligne des marques françaises, de l'Union européenne et internationales)
  - BOPI (Bulletins officiels de la propriété industrielle)

Il n'y a pas de Data Scientist en poste actuellement chez mon nouvel employeur, il n'y a pas non plus d'algorithme prédictif de machine learning pouvant aider au scoring, simplement un programme avec des règles empiriques sur la possibilité qu'une entreprise dépose son bilan dans les 12 mois qui viennent. C'est une des raisons qui m'ont décidé à travailler sur ce projet. En effet, être le premier à développer un nouveau système de calcul, faire partie des pionniers, plonger dans les innombrables variables à disposition afin de découvrir quelles sont les plus déterminantes à quelque chose d'excitant et grisant. Cela correspond aussi parfaitement à ce qu'un stage doit apporter et c'est un travail très

proche de ce que j'ai pu voir dans les divers projets que j'ai fait jusque-là dans le cadre de la formation Datat Scientist chez OpenClassrooms.

En arrivant dans l'entreprise j'ai donc un double travail à effectuer : tout d'abord une analyse des résultats actuels, à l'aide de statistiques exhaustives pour bien comprendre les tenants et les aboutissants du métier, commencer à isoler les variables les plus importantes, et trouver une méthode de notation qui permettra de comparer différents systèmes de notation. Ensuite la récupération de données dispersée dans des dizaines de bases et des centaines de tables, puis l'application d'algorithmes de machine learning pour enfin pouvoir comparer les résultats et aviser de la méthodologie possible à appliquer dans le futur.

3. Problématique

Dès le début de mon stage je m'aperçois que le développeur du programme utilisé actuellement ne travaille plus ici, et que personne ne connait de manière approfondie les règles internes qui gèrent la notation. C'est la raison pour laquelle il y a un besoin très fort d'analyse de l'existant en tout premier lieu.

Il est en effet primordial d'être bien conscient des forces et faiblesses du système en cours afin de savoir quelle partie est à améliorer en priorité. Répondre à des questions telles que : Quel est le pourcentage d'erreur note par note ? Y a-t-il des entreprises très bien notée qui finissent radiées ? Si oui pourquoi ? Ou bien quelles sont les sociétés qui s'en sortent bien malgré une note vraiment basse ? Et quels sont les éléments qui ont fait baissé leur note ? …

Comme rien n'a encore été fait je prends pleinement conscience du travail qu'il y a devant moi et commence par me plonger dans les bases de données afin de me familiariser avec et sortir quelques statistiques utiles pour la suite.

4. Missions confiées

A peu près à égale importance, j'ai donc deux missions à réaliser, la première étant un état des lieux sous forme statistique des données actuellement utilisées, et la seconde d'essayer différents algorithme de machine learning pour améliorer le système actuel de notation.

4.1. Mission 1 : Statistiques

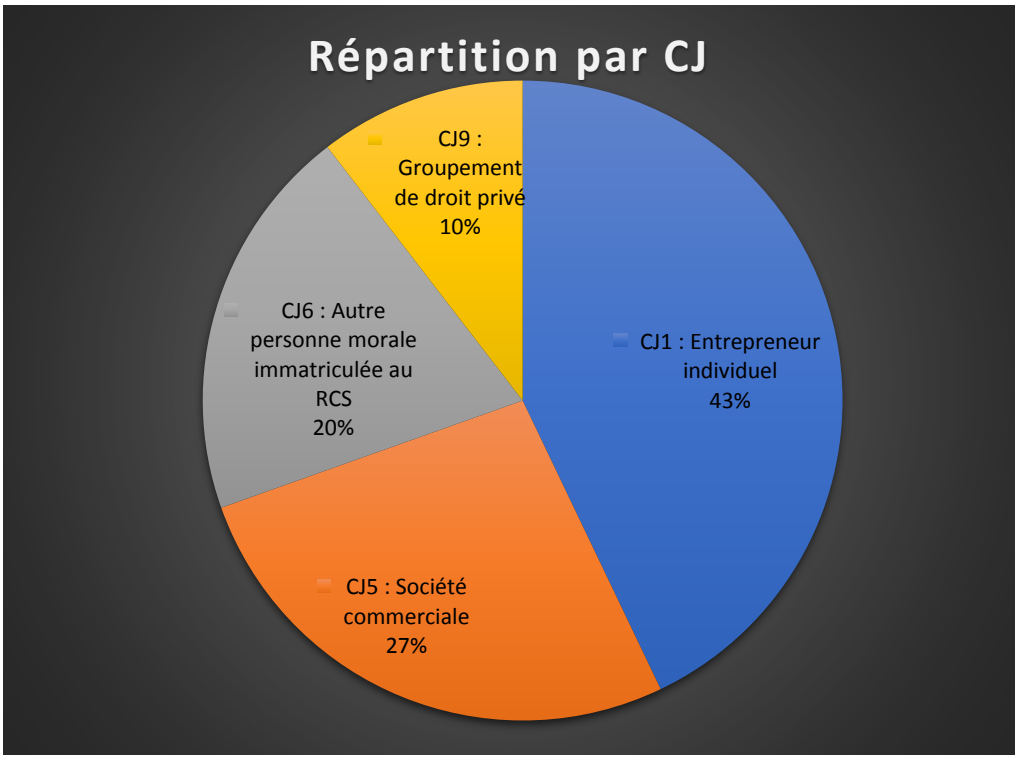
Nous allons donc plonger dans les bases de données de la société, c'est du MySQL, il y a 31 bases de données et des centaines de tables. La taille totale de l'ensemble fait plus de 900 Go, attention donc à ne pas se perdre dans la somme considérable d'information à notre disposition.

Commençons par lister les entreprises par catégorie juridique :

Catégories juridiques	Nb Siren	%
null	8 871	0,08
0	4 856	0,04
1 Entrepreneur individuel	4 537 267	41,54
2 Groupement de droit privé non doté de la personnalité morale	103 827	0,95
3 Personne morale de droit étranger	83 495	0,76
4 Personne morale de droit public soumise au droit commercial	1 697	0,02
5 Société commerciale	2 820 242	25,82
6 Autre personne morale immatriculée au RCS	2 114 154	19,35
7 Personne morale et organisme soumis au droit administratif	122 203	1,12
8 Organisme privé spécialisé	21 750	0,20
9 Groupement de droit privé	1 104 986	10,12
	10 923 348	100

On constate plusieurs choses intéressantes : Tout d'abord la répartition inégale des SIREN au sein des différents groupes, sur les neuf catégories juridiques existantes actuellement au niveau INSEE certaines ont à peine la population suffisante pour faire des statistiques. Attention aussi à certain SIREN qui n'ont même pas renseigné leur catégorie juridique.

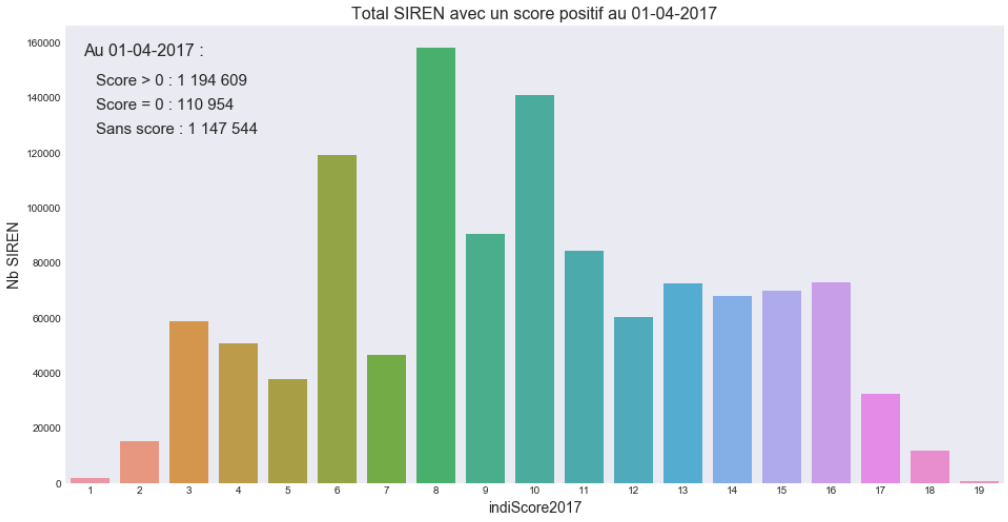
L'analyse brute et la modélisation de l'ensemble des SIREN risque d'être fastidieuse et compliquée, d'une part les 11 millions de SIREN constituent un corpus vraiment imposant, et d'autre part les règles de gestion des différentes catégories juridiques n'ont pas forcément de rapport entre elles. Il semble donc judicieux de faire une séparation à ce niveau.



Parmi les presque 11 millions de SIREN existants, on compte 2.8 millions de sociétés commerciales, soit à peu près un quart de la population.

Comme en plus la demande des clients se porte essentiellement sur l'analyses des sociétés commerciales qui constituent le cœur de métier, j'ai consacré le reste de mon stage sur ces dernières, la catégorie juridique 5, ou CJ5.

Intéressons-nous maintenant à la notation de ces entreprises, le score donné correspond à une note entre 0 et 20 dont on peut étudier la répartition. Regardons ce que ça donne avec les scores donnés il y a un an :

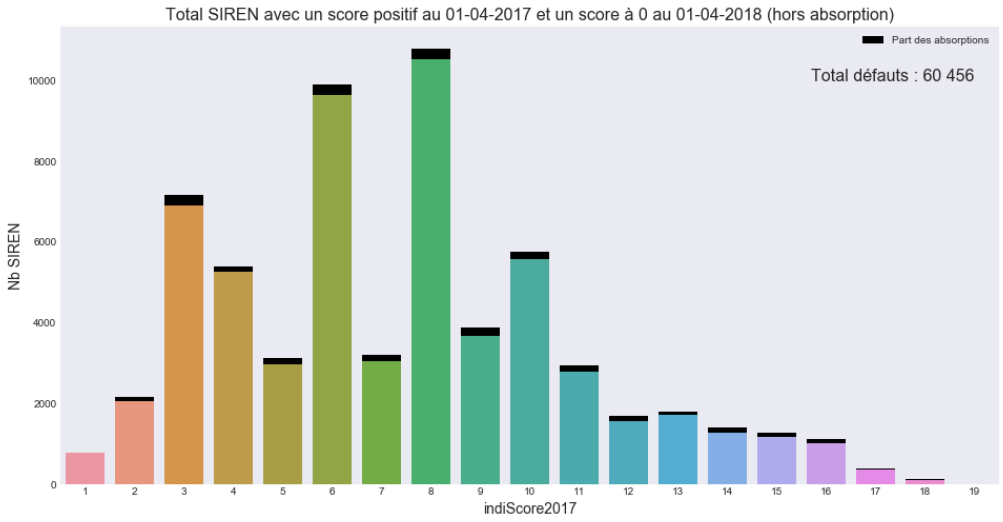


Comme on peut le voir sur le graphique ci-dessus les notes ne sont pas distribuées selon une loi statistique mais de manière empirique. Une partie importante semble être apportée par une note sur

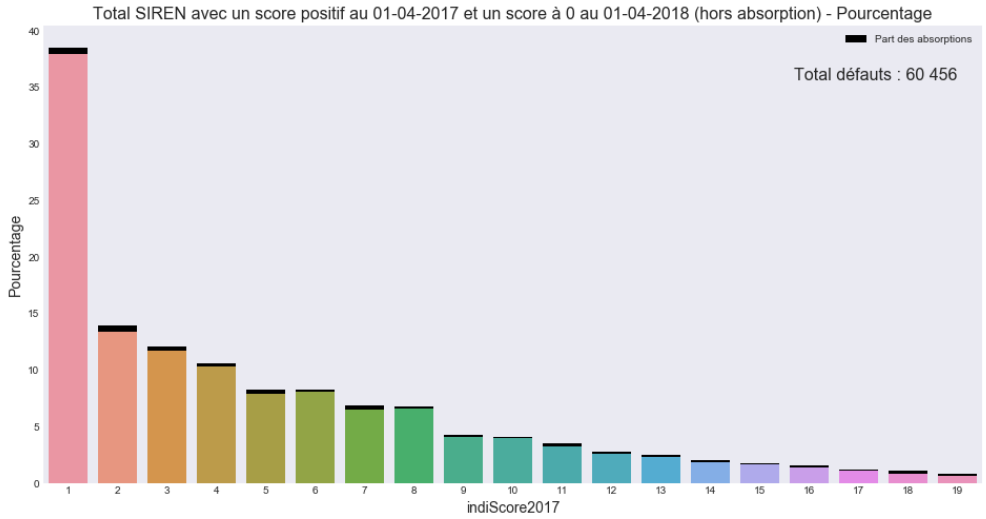
10 qui aurait été doublée pour faire une note sur 20, d'où le nombre important de sociétés avec les notes paires 6, 8 et 10.

Le score d'une entreprise donne une indication de sa santé financière et de sa capacité à être encore active à l'horizon d'un an. On peut résumer simplement en disant qu'une note de 6 ou moins désigne une entreprise qui est en difficulté et a un fort risque de défaillance dans les 12 mois qui viennent.

Maintenant que l'on a les notes des entreprises il y a 12 mois, on va pouvoir comparer avec l'état des lieus actuel. Donc si l'on regarde ces mêmes entreprises maintenant on va pouvoir dire si les notes de l'année dernière sont cohérentes. Regardons parmi les entreprises en défaut quelles étaient leur note 12 mois avant :



Ici une première analyse a montré que les entreprises absorbées (ou rachetées) tombaient aussi à 0, on va donc en tenir compte pour ne pas fausser les stats. On voit quand même que les entreprises qui ont fait faillite possédaient toutes les notes possibles 1 an avant. Regardons ce que ça donne en termes de pourcentages :



Bon la forme du graphique obtenu a la bonne tendance mais les pourcentages devraient idéalement être bien plus importants au début (pour les notes inférieures à 6) et carrément nuls au-dessus de 15 ou 16. En effet, les clients ont du mal à accepter qu'on donne une note de 18 à une entreprise qui va disparaitre dans l'année, ce genre de cas porte un coup à la crédibilité de la note et est préjudiciable pour les affaires.

Cette conclusion a pu être généralisée après analyse des années antérieures : je suis remonté jusqu'en 2014, et même si le nombre de données baisse drastiquement, la tendance reste la même et prouve qu'il faut absolument, d'une part optimiser les notes basses aux entreprises vraiment en difficulté, et d'autre part améliorer la notation afin d'éviter les cas extrêmes des entreprises très bien notées qui disparaissent finalement au cours de l'année.

4.2. Mission 2 : Modélisation

Les bases de données constituent un ensemble des plusieurs centaines de tables pour un poids total de plus de 900 Go avec 11 millions de SIREN différents. C'est la première difficulté du travail, il va falloir à défaut de tout prendre, sélectionner les bonnes tables et les bonnes variables qui pourront donner un premier résultat encourageant avant d'approfondir et d'aller plus loin avec la modélisation.

4.2.1. Analyse des données

Voici un aperçu des features récupérées :

- Venant de la table insee.identite:
  - Nature de l'activité
  - Modalité de l'activité
  - Exploitation de tous les moyens de production
  - Année de validité de l'activité principale
  - Catégorie juridique
  - Nombre d'établissement
  - Capital
  - Effectif
  - Tranche d'effectif
  - Code NAF
  - Département
  - Tranche de chiffre d'affaire
  - Tranche de chiffre d'affaire à l'export
- Venant d'autres tables :
  - Nombre de participation
  - Nombre d'actionnaires
  - Nombre de Personne Moral et Personne Physique
  - Cotation en bourse
  - Nombre de contentieux en tant que défenseur ou demandeur
  - Nombre de marques déposées
  - Valeurs principales du bilan, plus de 20 valeurs (avec historique sur 3 ans)
  - 'Score' sur 5 années (Score actuel)
  - 'encours' sur 5 années (Calculé avec le score)
  - 'procol' sur 5 années (Target)

C'est évidemment une petite partie de ce qu'on pourrait récupérer, mais intuitivement ce sont les données les plus représentatives et facilement récupérables pour une première étude.

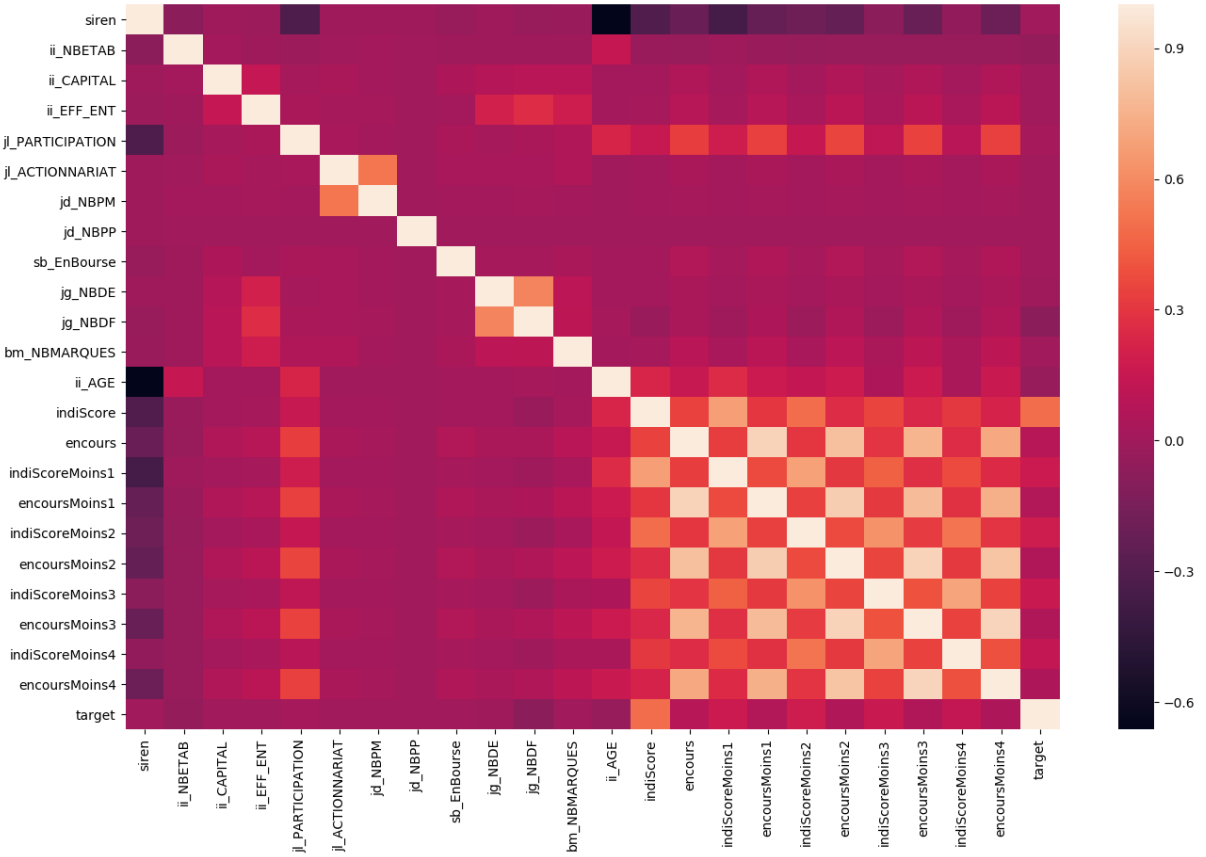
4.2.1.1. Les features numériques

Une description classique des données numériques va nous permettre d'avoir une idée de la distribution :

	siren	ii_NBETAB	ii_CAPITAL	ii_EFF_ENT	jl_PARTICIPATION	jl_ACTIONNARIAT	jd_NBPM	jd_NBPP
count	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06
mean	5.882103e+08	9.567984e-02	2.532139e+05	7.699615e+00	1.536431e-01	8.785798e-02	3.227594e-01	1.748487e-06
std	1.831789e+08	4.532915e-01	1.865733e+07	2.582966e+02	3.659322e-01	5.070686e+00	1.878656e+01	1.322303e-03
min	5.420120e+06	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	4.394304e+08	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	5.177029e+08	0.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
75%	8.042120e+08	0.000000e+00	5.000000e+03	3.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
max	9.999907e+08	1.660000e+02	7.441822e+09	2.200000e+05	3.000000e+00	3.720000e+03	7.769000e+03	1.000000e+00

sb_EnBourse	jg_NBDE	jg_NBDF	bm_NBMARQUES	ii_AGE	indiScore	encours
1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06	1.143846e+06
4.143914e-04	8.283807e-02	1.005258e-01	2.185347e-01	5.770733e+00	9.715782e+00	2.185649e+04
2.035240e-02	4.809504e+00	1.545537e+00	5.861196e+00	3.518753e+00	4.605785e+00	5.953284e+04
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.000000e+00	6.000000e+00	7.000000e+02
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	5.000000e+00	1.000000e+01	3.127000e+03
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+01	1.300000e+01	1.500000e+04
1.000000e+00	2.980000e+03	7.940000e+02	3.381000e+03	1.010000e+02	1.900000e+01	5.000000e+05

Voyons ce que donne la matrice de corrélation de Pearson :



Plusieurs choses ressortent de cette matrice de corrélation :

SIREN corrélé à ii_AGE : c'est logique, les nouvelles sociétés ont des numéro de SIREN incrémenté, et donc plus le SIREN est petit plus la société est ancienne.

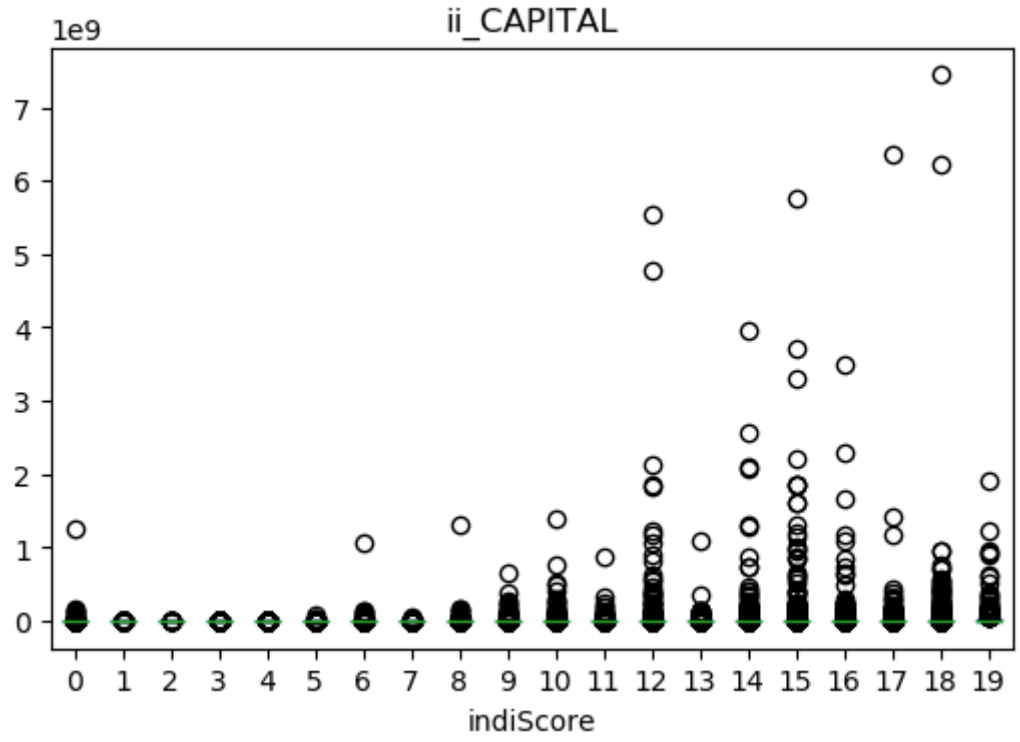
jl_ACTIONNARIAT corrélé à jd_NBPM : ici aussi c'est logique, le nombre de dirigeants en tant que personne morale est lié au nombre d'actionnaire.

jg_NBDE corrélé à jg_NBDF : les nombres de contentieux en tant que demandeur ou défenseur sont liés, on peut imaginer que si un client réclame un préjudice la société peut se retourner contre un prestataire par exemple, donc ça fera un contentieux en tant que défenseur et un en tant que demandeur.

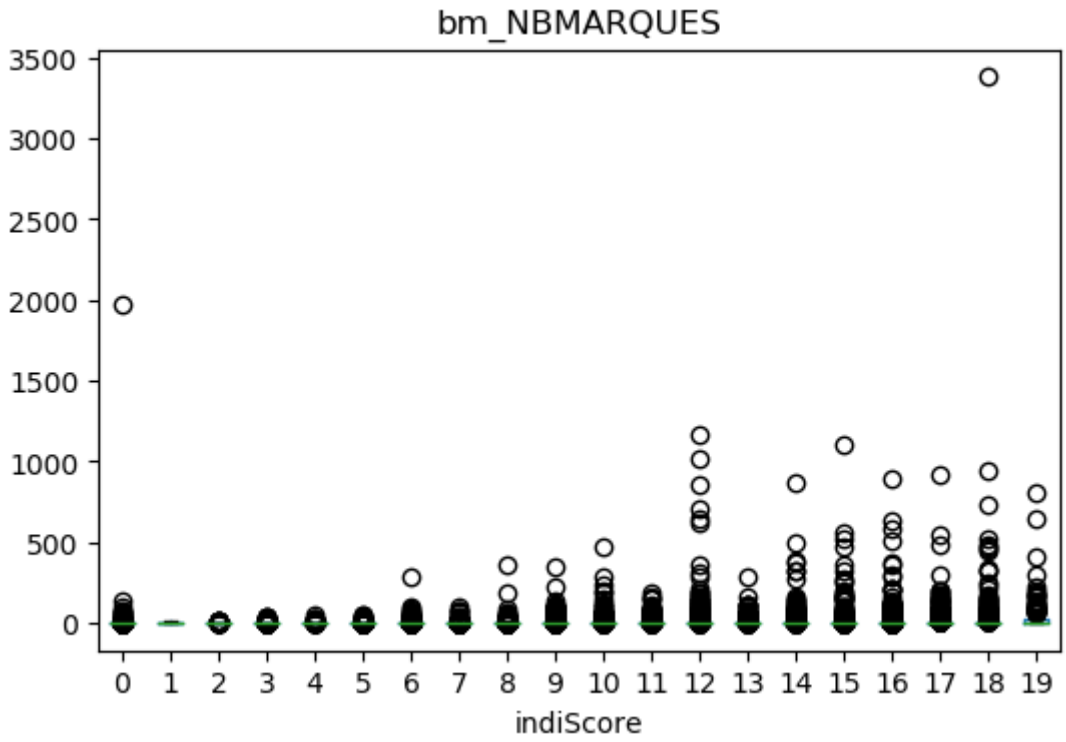
Les scores sont corrélés entre eux d'une année à l'autre, ça montre la relative stabilité dans le temps des notations, une entreprise bien notée aura tendance à le rester et pareil pour les entreprises mal notées. Ceci est dû au fait que les paramètres changent peu en moyenne d'une année à l'autre.

Les encours sont corrélés entre eux d'une année à l'autre pour les mêmes raisons que les scores.

On peut aller plus dans le détail sur les features qui nous intéresse à l'aide de quelques boites à moustache :



Pour le capital par exemple on peut voir que plus une société en possède, meilleure va être sa note.



Pour le nombre de marque c'est le même principe, on peut dire qu'une société qui possède beaucoup de marques déposée devrait être plus stable que les autres, c'est globalement ce que l'on voit mis à part quelques outliers.

D'ailleurs ces graphiques nous indiquent la façon de traiter les outliers et les valeurs manquantes : on peut soit borner les outliers, soit faire des buckets sur la feature (ce qui est déjà fait pour certaine, on a en effet des tranches d'effectifs ou des tranches de chiffre d'affaire).

4.2.1.2. Les features catégorielles

La description des features catégorielles ci-dessous ne montre pas de problème particulier, sauf pour ii_APE_ENT qui possède plus de 700 possibilités différentes et que l'on va traiter plus bas.

	ii_ACTIVNAT	ii_ORIGINE	ii_MODET	ii_EXPLET	ii_CJ	ii_APE_ENT	ii_TEFF_ENT	ii_ADR_DEP	ii_TCA	ii_TCAEXP	ii_NAF1	procol
count	2453107	2401741	2453107	465655	2453107	2453107	2453107	2453105	2453107	2453107	2453107	330429
unique	18	14	1	3	85	723	16	99	11	1	19	9
top	0	1	nan	O	5499	7022Z	00	75	nan	nan	G	P
freq	1311825	1884439	2453107	425241	1135772	108388	1169281	305408	2363855	2453107	526619	143738

Les départements

```
# Les départements les plus représentés
df.groupby('ii_ADR_DEP')['siren'].count().sort_values(ascending=False)[:10]
```

```
ii_ADR_DEP
75    120177
13     42826
92     42282
69     41432
93     36669
97     33077
59     32494
06     30045
33     28391
94     25125
```

On peut voir que Paris est très largement plus représenté que n'importe quel autre département avec 120 000 entreprises commerciales recensées. Suivent de loin mais groupés les départements correspondant aux grandes villes comme Marseille, Lyon, Lille, Nice et Bordeaux ainsi que les départements de la région parisienne (92, 93, 94), et on a aussi le groupe des DOM (97) très bien fourni en entreprises commerciales.

Les catégories juridiques

```
# Les catégories juridiques les plus représentées
df.groupby('ii_CJ')['siren'].count().sort_values(ascending=False)[:5]
```

```
ii_CJ
5499    514320
5710    211786
5498    207856
5720    146445
5202     17704
```

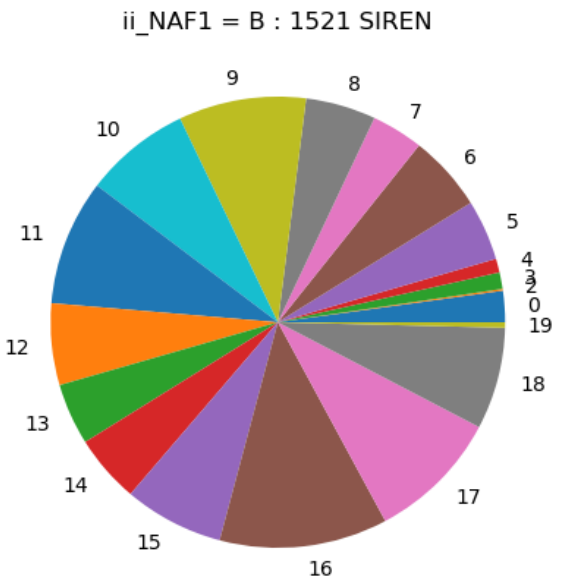
On peut remarquer que la catégorie juridique la plus représentée a pour code 5499 qui correspond aux SARL. Ensuite viennent les SAS (5710), les SARL unipersonnelles (5498), les SAS unipersonnelles (5720) et les sociétés en nom collectif (5202).

Le code NAF

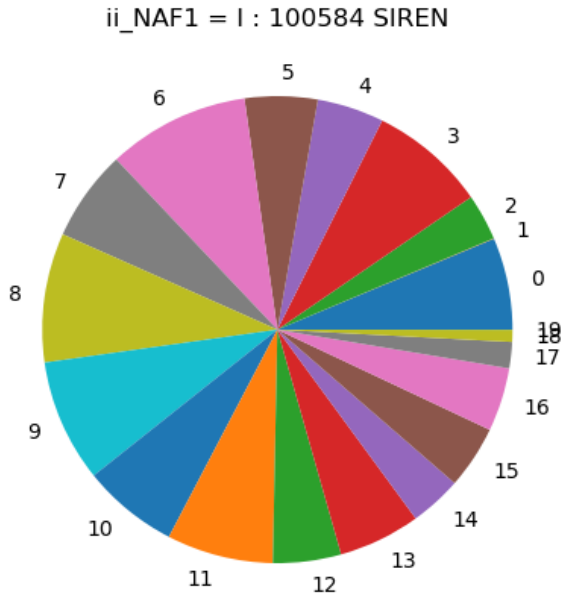
Le champ ii_APE_ENT correspond au code NAF de niveau 5. On en trouve plus de 700 occurrences différentes en base de données, ce qui est beaucoup pour une variable catégorielle. Par chance il est possible de le simplifier jusqu'au code NAF de niveau 1 qui ne possède plus que 19 cas possibles. C'est ce qu'on a fait avec ii_NAF1.

Examinons un peu la répartition des notes selon le code NAF à travers deux exemples :

NAF B : Industries extractives



NAF I : Hébergement et restauration



On voit bien grâce à ces graphiques qu'en moyenne les hôtels et restaurants sont beaucoup moins bien notés que les gros industriels. C'est sûr que le modèle économique n'a rien à voir, le secteur du tourisme est plus volatile, les sommes engagées pour monter un point de restauration n'ont rien à voir

avec le coût d'une usine et tout le monde peut tenter de monter son propre restaurant. Le taux de faillite est donc sans commune mesure, ce qui explique que les notes sont bien plus basses et que la population du secteur est beaucoup plus grande.

La feature semble donc très importante, peut-être au point de séparer les modèles, ça pourra être une des pistes d'amélioration dans le futur.

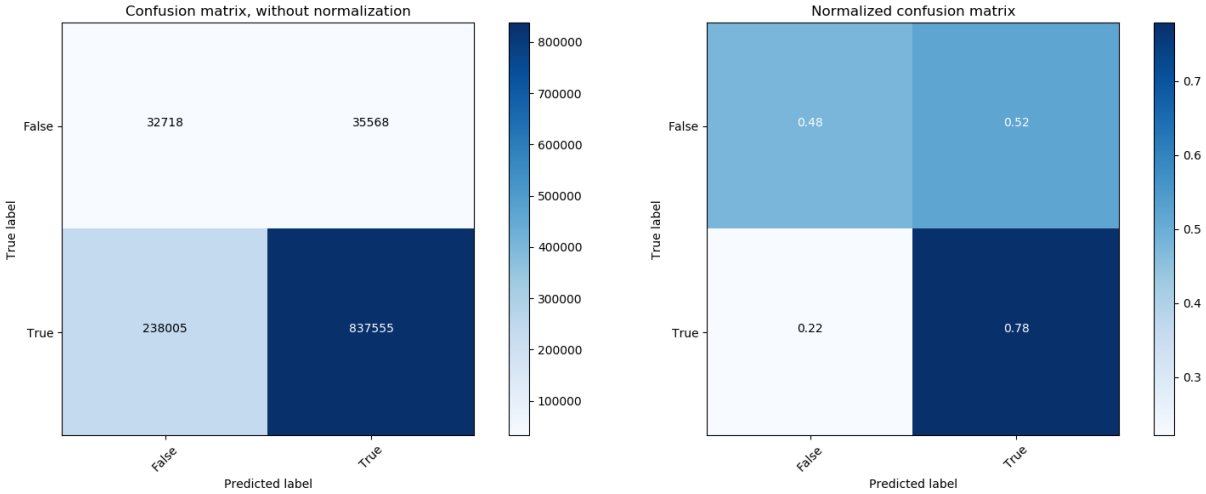
4.2.1.3. La target

Alors voilà la question peut-être la plus importante : à quoi correspond la target ?

Si le champ 'procol' indique une procédure collective, c'est le début d'une radiation de la société et c'est ce qui déclenche un score à 0. C'est donc notre target.

Connaissant la prédiction d'il y a un an, on peut construire une matrice de confusion représentant le modèle actuel :

- Note <= 6 il y a un an : Prédiction Fausse
- Note > 6 il y a un an : Prédiction Vrai
- Pas de procédure collective en cours : Target Vrai
- Procédure collective en cours : Target Fausse




```
AUC = 0.628
Spécificité = 0.478
Precision = 0.959
Recall = 0.779
```

On retrouve ce que l'on a commencé à analyser plus haut : il y a beaucoup trop de faux négatifs (près de 240000, ce qui est énorme même si le pourcentage n'est pas si horrible), et surtout il y a un taux de faux positifs bien trop important, plus de la moitié des entreprises en faillite n'ont pas été repérées.

Donc même si la précision est bonne on va chercher à améliorer la spécificité et le recall (la sensibilité).

4.2.2. Nettoyage et préparation

La phase de nettoyage comporte :

- La suppression des colonnes sans valeurs
- Un typage non numérique pour les feature catégorielles, pour un traitement automatisé juste après
- Quelques transformation : NAF de niveau 5 en NAF de niveau 1, date de création en âge…
- Correction de la target en prenant les absorptions en compte
- Remplissage des valeurs manquantes
- Bornage des outliers

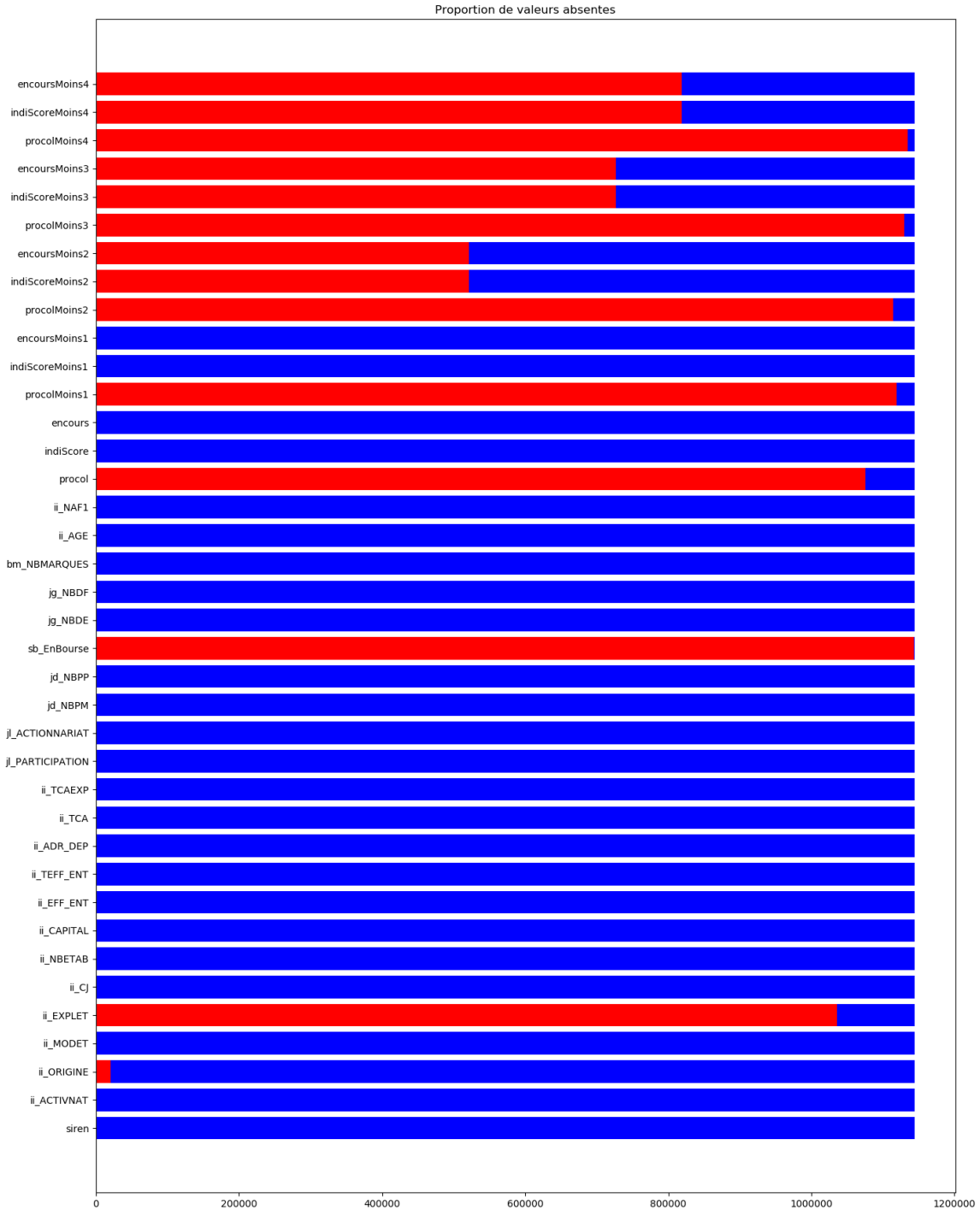
4.2.2.1. Les outliers

Pour définir un outlier je vais utiliser la méthode de Tukey en calculant les quartiles et distances inter quartiles, je vais en revanche borner mes valeurs à 10 IQR pour ne pas trop écraser les données, cap qui pourra être modifié par la suite en fonction des tests.

Comme on a fait un tri des features par leur type auparavant on peut utiliser la méthode de manière automatique en bouclant sur les colonne du dataframe.

Tous ces traitements sont écrits en Python sous forme de fonctions paramétrables afin de pourvoir rejouer l'analyse et la modélisation facilement en changeant juste les paramètres désirés.

4.2.2.2. Les valeurs absentes



Comme on peut le voir sur le graphique certaines features sont peu renseignées et possèdent un nombre très important de valeurs absentes.

En se basant sur le type de la colonne, sachant qu'il a été éventuellement modifié en toute connaissance de cause, on peut trier les features catégorielles et ainsi remplacer les valeurs absentes par la valeur la plus courante.

Ensuite pour les features numériques on remplace les valeurs manquantes par la médiane afin de limiter l'influence des outliers (même si normalement ils ont été bornés auparavant).

4.2.2.3. La mise à l'échelle

On utilise un outil standard, le StandardScaler pour mettre à l'échelle nos données numériques. Ce scaler est entrainé avec le jeu d'entrainement et enregistré pour de futures utilisation, aussi bien avec le jeu de test que pour les prédictions une fois que le modèle est défini.

4.2.2.4. La dummisation

Toujours en se basant sur le type des colonnes on peut transformer toutes nos données catégorielles avec un get_dummies.

Le gros avantage qu'il y a à se baser sur le type des colonnes est que l'on va pouvoir ajouter des features au fur et à mesure des explorations dans les bases de données, et le travail de préparation devrait continuer de se faire sans nouveau code, ce qui va accélérer grandement les tests pour la suite.

4.2.3. Classification binaire

Nous avons maintenant tout ce qu'il faut pour entrainer un algorithme de classification binaire.

Nous pouvons faire une recherche d'un modèle en utilisant l'AUC (l'aire sous la courbe ROC) comme notation pour comparer les différents tests, sachant que le programme actuel a une AUC de 0.628.

4.2.3.1. Cas 1 : Toutes les variables

Dans un premier temps on va utiliser toutes les variables à notre disposition pour se donner un maximum de chance de faire un meilleur score que le système actuel.

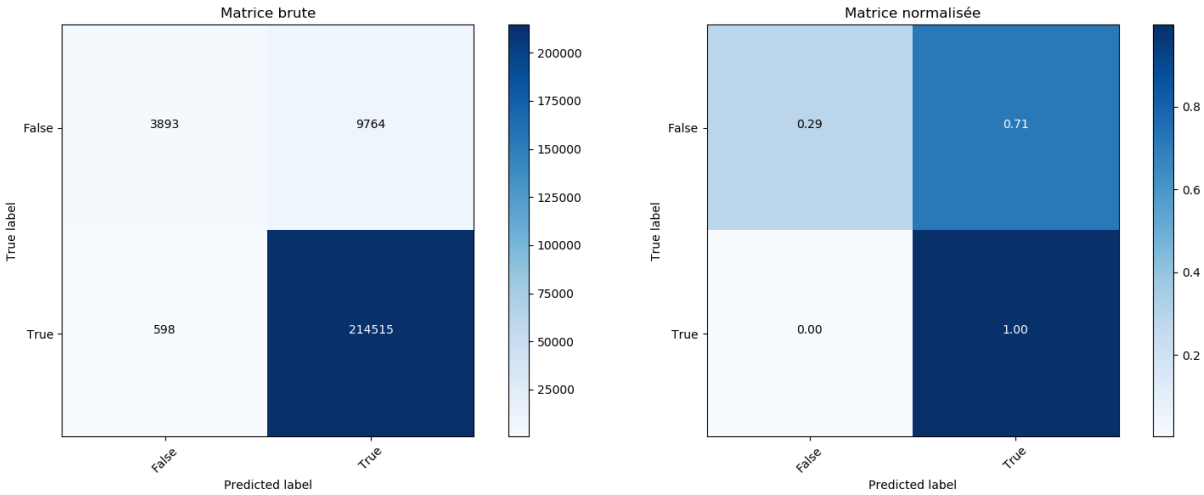
Essayons avec un panel d'algorithmes parmi les plus classiques :

- Classifieur naif
  - AUC = 0.5 comme prévu
- Gaussian Naive Bayes
  - Pas performant, AUC faible et calcul long
- KNN
  - Calcul vraiment trop long
- Logistic Regression
- Extra Trees
- Random Forest
- Grandient Boosting
  - Calcul vraiment trop long

Les tests opérés sur un échantillon permettent une première sélection de modèles et je vais garder la régression logistique, les extra trees et la random forest pour faire des tests plus poussés.

4.2.3.1.1. Logistic Regression

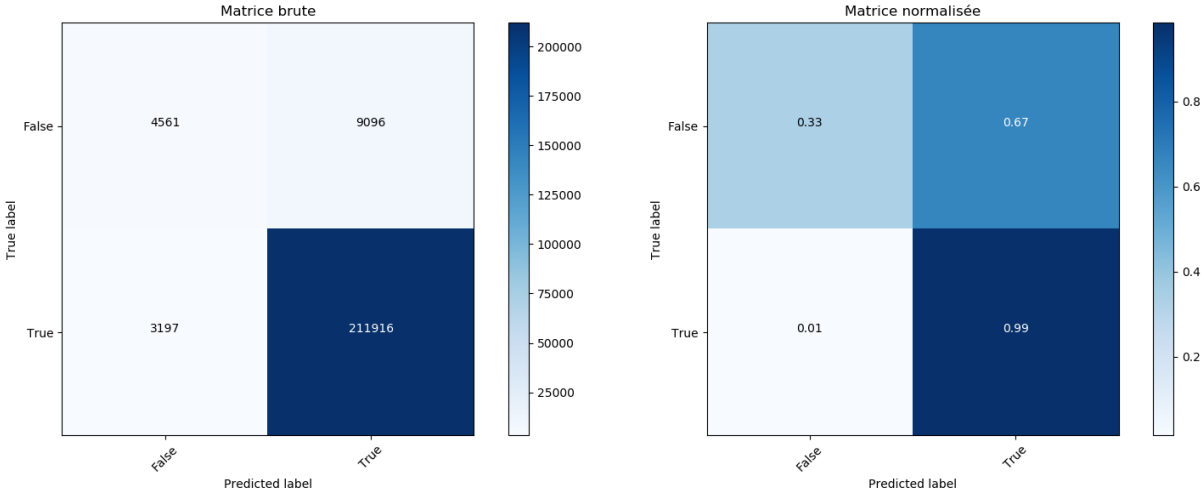
```
AUC = 0.820
Spécificité = 0.285
Precision = 0.956
Recall = 0.997
```



La régression logistique donne vraiment un très bon résultat en terme d'AUC et de recall, en revanche et c'est peut-être le plus important la spécificité reste haute ce qui n'est pas un bon point pour la suite.

4.2.3.1.2. ExtraTrees

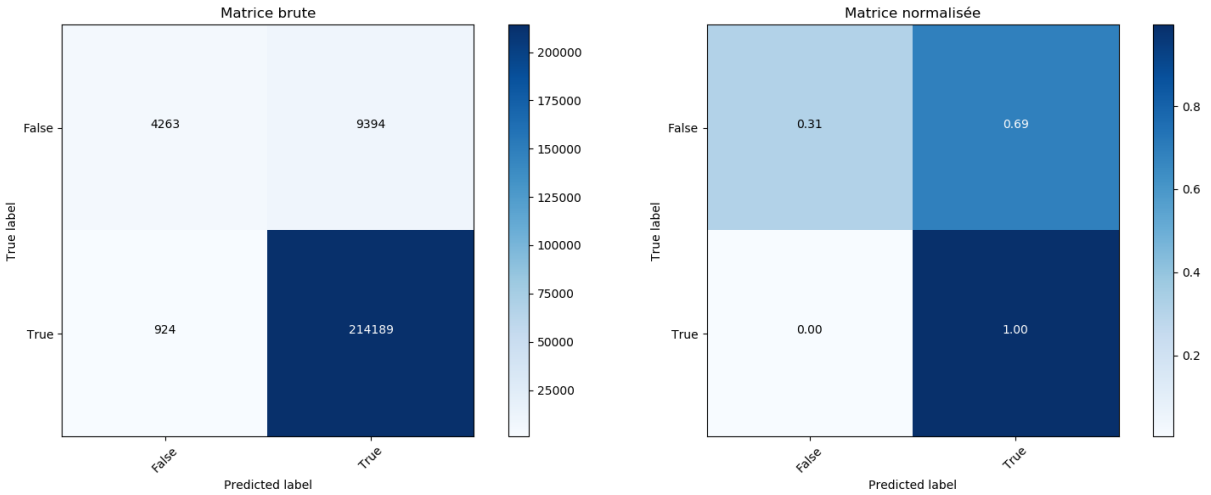
```
AUC = 0.740
Spécificité = 0.334
Precision = 0.959
Recall = 0.985
```



Les extra trees montrent eux une AUC bien moins bonne mais un taux de faux positifs bien meilleur.

4.2.3.1.3. Random Forest

```
AUC = 0.816
Spécificité = 0.312
Precision = 0.958
Recall = 0.996
```



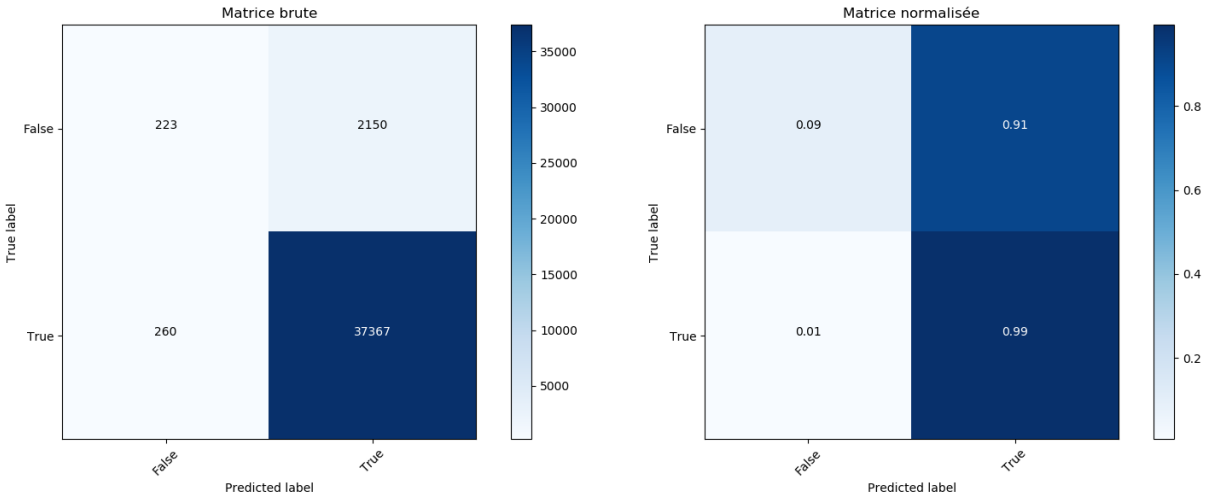
Avec la random forest on a un bon compromis entre les modèles précédents, une bonne AUC et une bonne sensibilité (recall). Ça plus les capacités du modèle à expliquer ses choix de manière fonctionnelle font que ce sont les random forets qui sont retenus pour la suite du projet.

4.2.3.2. Cas 2 : features basiques uniquement

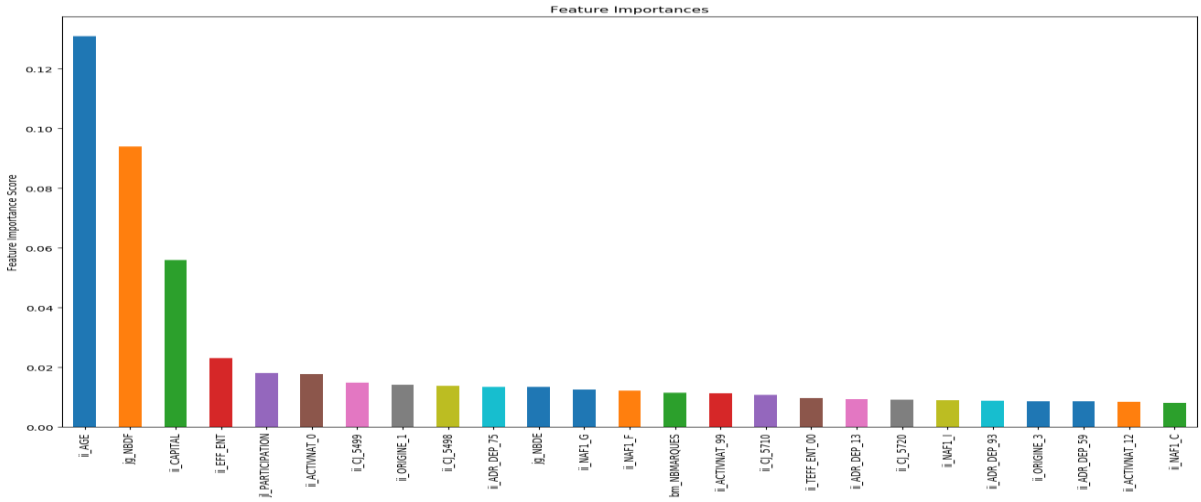
Dans un second temps je vais garder uniquement les features brutes qu'on peut récupérer parmi les bases de données à notre disposition. Ce test a pour but de découvrir éventuellement des variables cachées et influentes pour la bonne santé d'une société, c'est pour cela qu'on enlève tout ce qui résulte d'une pré analyse utilisant déjà les données de base : les features 'maison' score et encours.

- Random Forest pour comparer l'importance des variables

Le cas 2 est étudié pour constater l'importance relatives des features basiques dans la santé financière d'une entreprise, en enlevant donc les données résultant déjà d'un calcul interne.



Les résultats, surtout en terme de détection des faux positifs sont bien moins bons, on ne retiendra donc pas ce modèle pour la finalité du projet mais on va surtout s'en servir pour analyser l'importance relative des features entre elles.



L'âge d'une entreprise semble donc être la variable la plus importante parmi celles étudiées pour déterminer si une entreprise va bien ou pas. C'est effectivement ce dont se doutait mon maitre de stage qui connait bien le métier, mais jusqu'à présent il ne pouvait pas le prouver et donc pouvait difficilement s'en servir comme justificatif devant un client. Plus une entreprise est âgée, plus elle est armée pour tenir face à ses concurrents. Au moins cette étude aura permis de lever ce genre de doute et d'autoriser la diffusion de ce genre d'informations.

De même on constate l'importance du nombre de contentieux en tant que défenseur. On imagine qu'une entreprise attaquée de toute part par des créanciers ou des clients a des soucis à se faire pour la suite de son existence.

Le capital vient en troisième position. De manière logique, une société avec un fort capital aura plus de chance de passer l'année sans encombre.

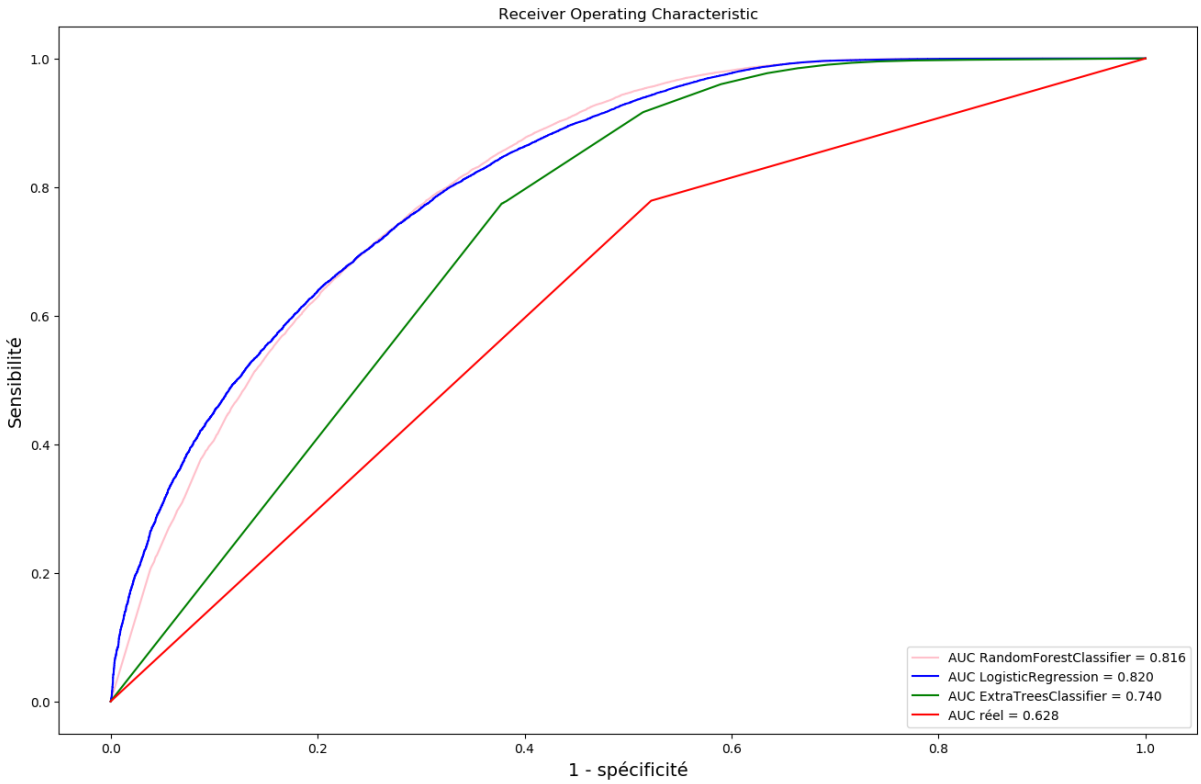
Les effectifs de l'entreprise viennent ensuite, puis différentes autres caractéristiques avec de moins en moins d'importance.

Il va sans dire que tous ces résultats devront quand même être vérifiés régulièrement et peut-être affinés en ciblant la population de manière plus précise et dans différents cas de figure.

Il faut aussi se méfier du taux de remplissage de toutes ces features qui parfois ne sont pas obligatoires et donc moins fiables.

4.2.4. Choix final du modèle

Afin de comparer nos différents résultats affichons les courbes ROC sur un même graphique :

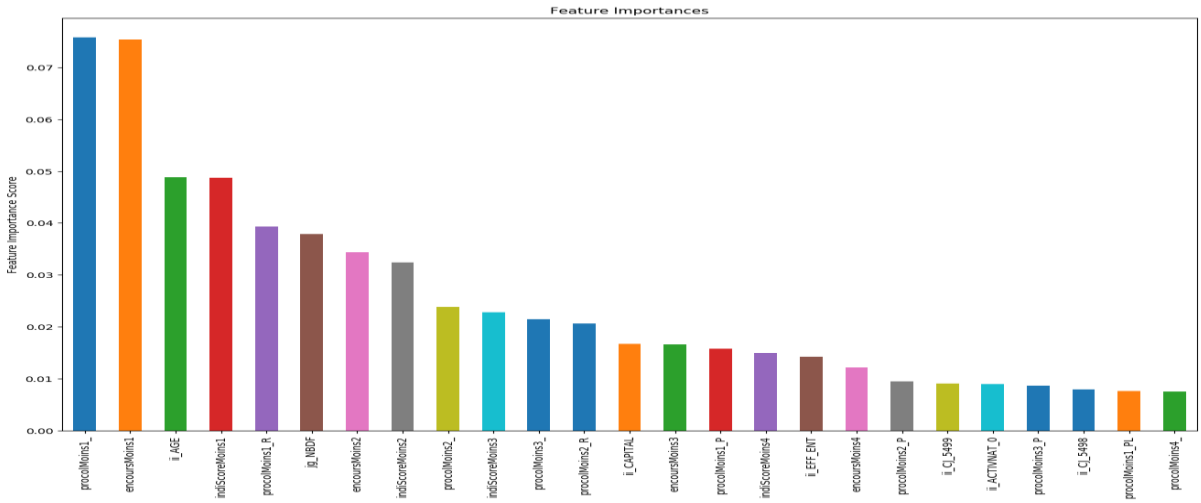


Ce tableau comparatif peut aussi nous guider dans le choix de notre modèle :

	Temps de calcul (s)	Taille du modèle	Score (AUC)
RandomForestClassifier	3900	1,2 Go	0,816
LogisticRegression	53000	4 Ko	0,820
ExtraTreesClassifier	660	267 Mo	0,740

Effectivement, pour une AUC équivalente, et sachant que la taille du modèle n'est pas une contrainte contrairement au temps de calcul, la random forest l'emporte sur la régression logistique, d'autant que son recall est aussi meilleur.

Voyons ce que donne le graphique des features les plus importantes :



Par rapport au précédent graphique avec seulement les features basiques dont l'ordre d'importance est conservé fort heureusement, on voit apparaitre en top position les scores et encours calculés l'année précédente. Ce sont des informations à considérer avec précaution car il n'y a plus de connaissance en interne pour expliquer en détail l'algorithme de programmation qui a abouti à ces résultats. Mais en même temps le fait d'avoir utilisé un nombre important de variables relatives à l'entreprise mais aussi à son secteur d'activité ou à l'historique de ses dirigeants fourni des informations, certes compilées, mais d'une énorme importance comme le prouve ce graphique.

Tant qu'on n'aura pas extrait une à une les mêmes informations que celles utilisées par le programme de calcul actuel il sera très utile de garder ses résultats en tant que feature d'entrée. Même si tous les détails sont écrasés et remplacés par une note finale, cette dernière reste quand une indication forte de ce qu'on cherche.

4.2.5. Optimisation du modèle

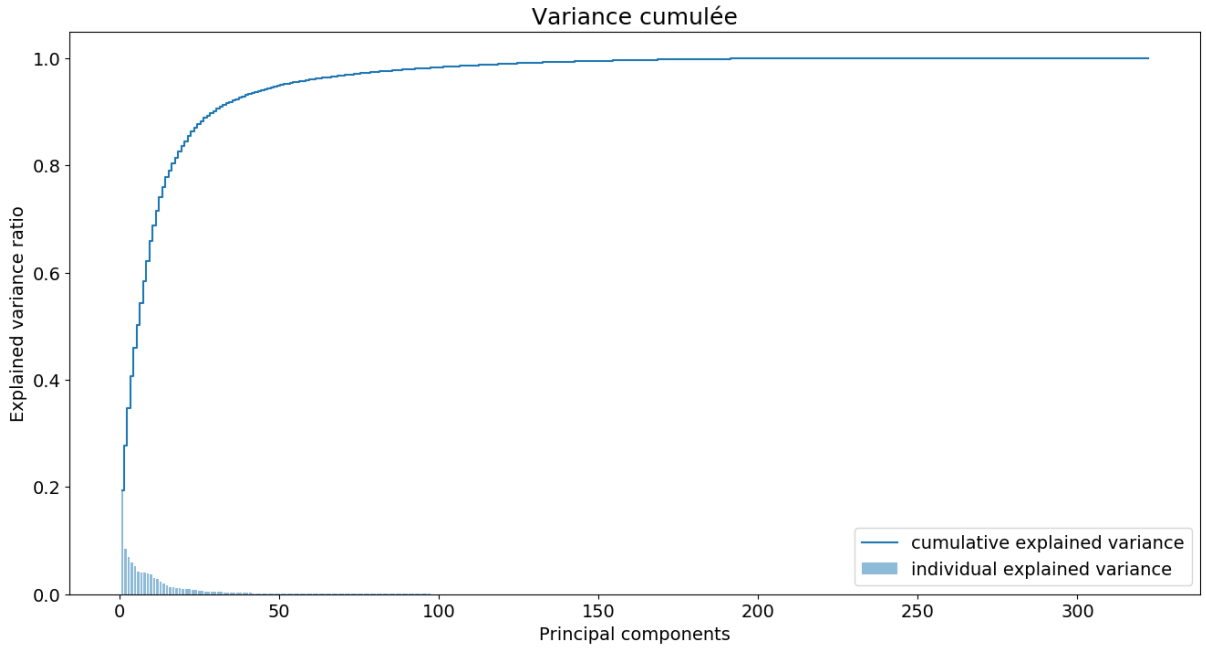
4.2.5.1. Les dimensions

Le nombre de dimensions de notre corpus peut s'élever rapidement avec l'ajout de nouvelles features. Actuellement on en est au minimum à 320 dimensions, et ce en ayant déjà fait l'impasse sur les données financières (au moins dans un premier temps car tous les SIREN n'en possède pas loin de là) et en ayant simplifié le code NAF qui potentiellement peut ajouter 700 dimensions.

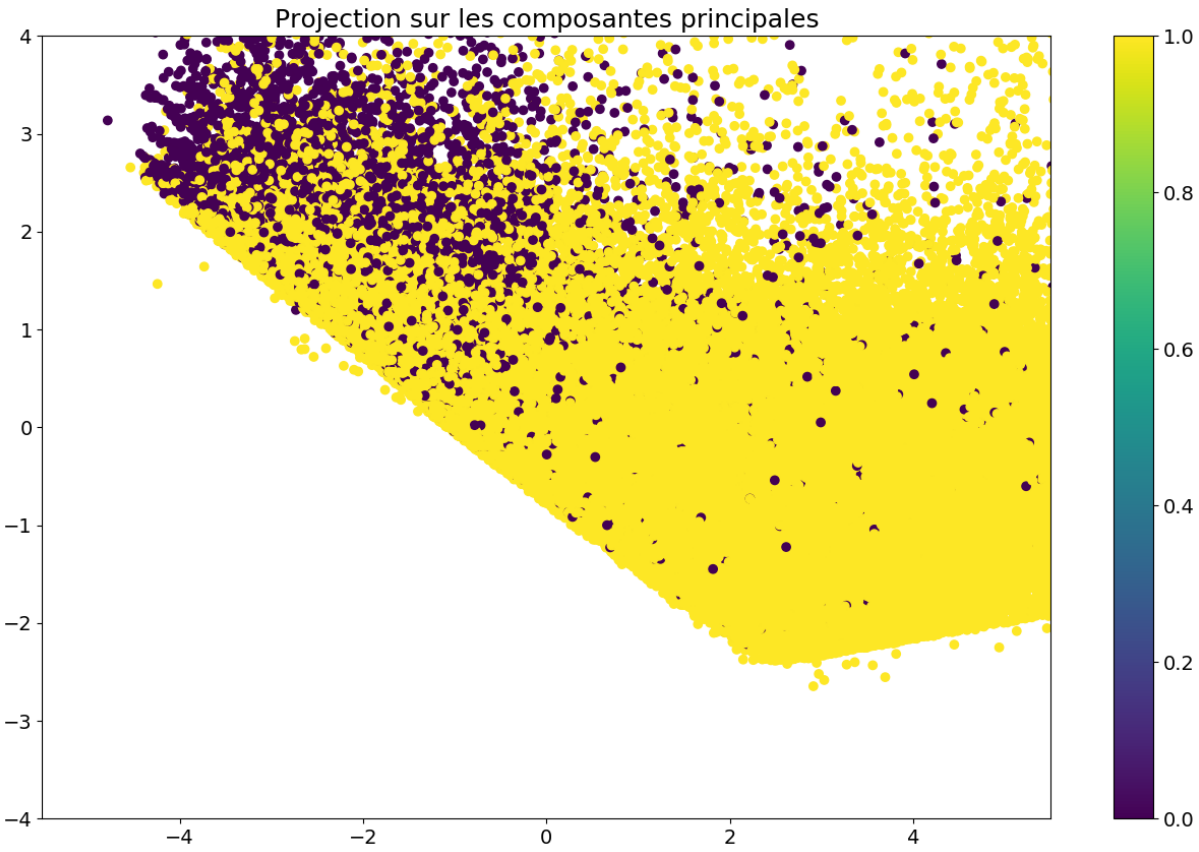
J'ai donc effectué des tests de réduction dimensionnelle à l'aide d'une ACP et d'une TSVD. Les résultats sont identiques avec les deux méthodes : 12 dimensions seulement permettent d'expliquer 90% de la variance (voir le graphique sur la variance cumulée ci-dessous).

C'est un résultat très intéressant car on sait qu'on va pouvoir accélérer les calculs si on le souhaite pour continuer les recherches et comparer d'autres modèles ou paramétrages.

Il faut garder cette possibilité en cas de besoin, mais pour notre projet il est important de pouvoir sortir la liste des features avec leur importance relative car il y a une forte demande de compréhension métier pour les fonctionnels, donc on va poursuivre avec une random forest sans réduction dimensionnelle.



La projection des données sur les deux composantes principales, qui expliquent 27.8% de la variance montre un début de segmentation mais on n'a pas encore assez de dimensions pour voir une séparation claire de nos deux groupes. C'est aussi normal car au final on recherche un système de notation continue, c'est-à-dire une régression et il n'y a pas en réalité deux groupes distincts d'entreprises mais tout un panel allant du très bien portant aux entreprises en grande difficulté financière.



4.2.5.2. Les paramètres

Essayons maintenant d'affiner les paramètres de notre random forest. On va utiliser un GridSearch avec une cross validation réglée à 5 dossiers et balayer une série de valeurs pour les paramètres les plus importants.

```
params = {
   'n_estimators': [100, 200, 500],
   'max_depth': [None, 1, 2, 3],
   'max_features': [None,'sqrt','log2','auto'],
   'max_leaf_nodes' :[None, 2, 3, 4],
   'min_samples_split': [2, 3]
}
```

Le résultat donne comme modèle optimal :

- n_estimator = 100:

- max_depth = None
- max_feature = log2
- max_leaf_nodes = 2
- min_sample_split = 3

4.2.6. Scoring

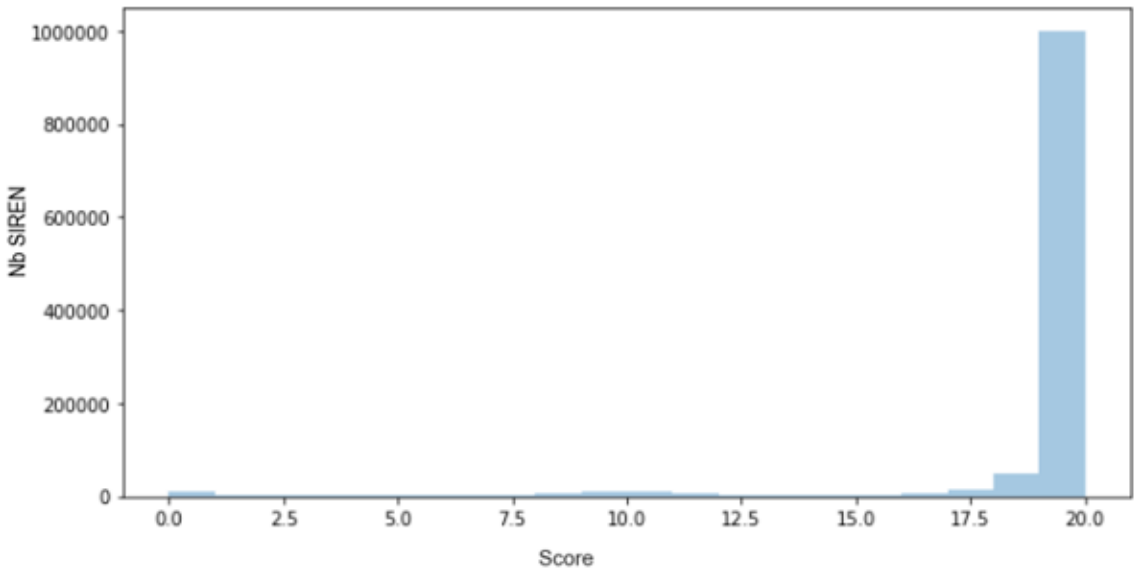
On peut maintenant utiliser le modèle qu'on vient de définir pour donner une note actuelle à nos entreprises. En effet, en utilisant les probabilités sur la prédiction que nous donne notre modèle il est aisé de rapporter les résultats sur une note sur 20.

On va donc pouvoir utiliser non seulement notre modèle sauvegardé mais aussi le scaler et la liste des catégories permettant la binarisation de nos données dans les mêmes dimensions que lors de l'entrainement de notre modèle.

Il faut aussi se rappeler que pour entrainer le modèle on s'est positionné 12 mois dans le passé et ainsi comparer notre prédiction avec l'état actuel des entreprise.

Ainsi pour effectuer une prédiction à 12 mois il va falloir adapter les features temporelles et faire le décalage adéquat pour s'intégrer à notre modèle.

Voyons la répartition des notes une fois la probabilité ramenée à une note sur 20 :



On constate une très grande majorité des notes proches de 20, c'est la transcription directe d'une forte probabilité de bonne santé de la majorité des entreprises. Il n'est pas si évident donc d'étaler le résultat d'une classification binaire sur une plage de note de 0 à 20. Mais en même temps cela reflète bien ce qu'on voulait, à savoir estimer si oui ou non une entreprise est assez solide pour passer l'année.

On voit aussi deux autres groupes de notes, les notes quasi nulles pour lesquelles on prédit peu de chance à l'entreprise, mais aussi il y a tout un groupe de notes qui oscillent autour de la moyenne, c'est ici qu'on a de la marge pour régler l'algorithme, le seuil qu'on va choisir peut nous permettre d'améliorer ou le taux de faux positifs ou celui des faux négatifs.

Il faudra voir avec les fonctionnels et peut-être aussi les clients si cette répartition leur va, sachant qu'on pourra toujours trouver un moyen d'étaler la répartition comme on veut.

5. Pistes d'amélioration

Le modèle obtenu obtient déjà de bons résultats dans certaines catégories (AUC, recall) mais doit encore être amélioré au niveau de la spécificité.

Pour cela il faut continuer d'extraire des variables potentiellement intéressantes de la base de données et les ajouter dans le dataset de départ. Le code que j'ai développé permet alors d'entrainer les modèles choisis sans modification et de pouvoir ainsi facilement comparer les résultats obtenus.

De la même manière il est possible d'augmenter l'historique des variables temporelles ou d'ajouter de nouveaux historiques de données en modifiant le code de manière minimale (il faut quand même indiquer quelle plage on étudie).

La signification du score actuel peut aussi être modifiée pour voir si son utilisation peut être améliorée. Par exemple si on indique une entreprise dans le rouge (alerte) sur un score inférieur à 5 ou 7 plutôt que 6 la lecture du score actuel peut en être amélioré.

6. Bilan

Ce stage chez m'a permis de découvrir tout un monde professionnel que je ne connaissais pas, sans parler d'un groupe de collègues admirables et attachants.

Bien sûr que je n'ai pas pu tout voir et qu'il reste encore un travail considérable pour aller au bout de l'utilisation des données libres sur les entreprises mais j'ai posé les bases d'un nouveau système de notation, de la récupération de données par script, à l'attribution d'une note finale représentant la capacité d'une entreprise à être encore active à l'horizon d'un an, en passant par l'analyse statistique de la situation actuelle et la recherche et l'optimisation d'un modèle prédictif.

Mieux que les bases, les résultats obtenus sont déjà meilleurs que le système actuel et certaines questions que se posaient mon maitre de stage ont déjà trouvé une réponse.

Je laisse un travail évolutif qui permet d'ajouter de nouvelles features à la volée pour enrichir le modèle et atteindre de meilleurs résultats sans modifier le code.

7. Synthèse

Lorsque l'on fait un travail, que ce soit de la recherche ou du développement, il faut toujours se demander ce qui va prendre le plus de temps : faire en boucle toute sorte de tests en notant les résultats sur un coin de cahier et essayer de synthétiser le tout à la fin ? ou bien développer dès le début un peu de code permettant d'automatiser au maximum la chaine d'actions à réaliser à chaque tests (récupération des données, nettoyage, calculs, analyse des résultats…) ?

Mon expérience m'a prouvé que l'automatisation dès le début du processus est toujours profitable, non seulement pour soi car ça permet d'avoir une meilleure vision globale du système, et de gagner énormément de temps lorsqu'on commence à industrialiser les tests. C'est aussi important pour les personnes qui auront à travailler autour du projet car le fait de pouvoir lancer des tests sans avoir à se pencher sur l'ensemble de la chaine d'action permet à plus de personnes de pouvoir faire des recherches et ainsi au final d'avoir plus de résultats et certainement une meilleure compréhension du modèle et du métier pour tout le monde.

Une fois cette méthodologie adoptée et que l'on peut lancer différents tests sur différents jeux de données simultanément, il est plus aisé de se concentrer sur la signification métier de chaque analyse et d'en tirer des conclusions avec plus de valeur ajoutée.

La recherche pure du meilleur résultat ne doit en effet pas masquer le fait que la compréhension de l'influence des données est extrêmement importante, car les résultats devront être compris et acceptés non seulement par les collaborateurs du projet mais aussi et surtout par les clients finaux qui ne manqueront pas de poser tout un tas de question sur le comment et le pourquoi de ce nouveau mode de calcul, que ce soit les notes d'une entreprise, les prévisions de comportement, les propositions de produits associés à un achat ou tout autre projet de machine learning.

8. Remerciements

Ce stage m'a donné plus confiance en moi dans le fait de pouvoir apporter une réelle valeur ajoutée dans une entreprise, même sans connaissance métier au préalable, et le tout en seulement quelques semaines.

Je me suis aussi rendu compte que le poste de Data Scientist, de par son rôle central entre les acteurs fonctionnels, techniques ou décisionnels et les données, est aussi une formidable opportunité pour créer du lien et parfaire ses connaissances métier.

C'est donc avec satisfaction que ce stage me conforte dans mon désir de poursuivre ma carrière en tant que Data Scientist, de préférence au sein d'une petite entité à taille humaine qui permet des décisions rapides et donc une réactivité accrue par rapport à une plus grosse structure.

Merci donc PDG de m'avoir accueilli dans son entreprise, de m'avoir fait confiance et mis à ma disposition tous les moyens techniques nécessaires au bon déroulement de ce stage. Merci aussi à mon maitre de stage, pour sa patience et ses conseils dans la compréhension de son métier et des enjeux professionnel qui en découlent.