Prise de notes sur modélisation

Planning :

|  |  |  |
| --- | --- | --- |
| * 28 mars |  | Avoir quasi fini la modélisation + l’analyse des données, l’avoir documenté dans un word   * EDA * Feature selection * Prépro * Essai de plusieurs modèles * Grid Search * Bagging / Boosting * MLFlow |
| * 25 avril => AGILE | Agile | - |
| * 23 mai => ascension + certif Agile sur 2 jours | Ascension | Faire la mise en DB et le packaging API (ie modules aussi en POO) ?  Revoir tuto sur streamlit |
| * 20 juin => 10 ans de mariage + certif azure ? | 10 ans de mariage |  |
| * 18 juillet |  |  |
| * 15 aout |  |  |

* 28 mars
* 25 avril => AGILE
* 23 mai => ascension + certif agile sur 2 jours
* 20 juin => 10 ans de mariage + certif azure ?
* 18 juillet
* 15 aout

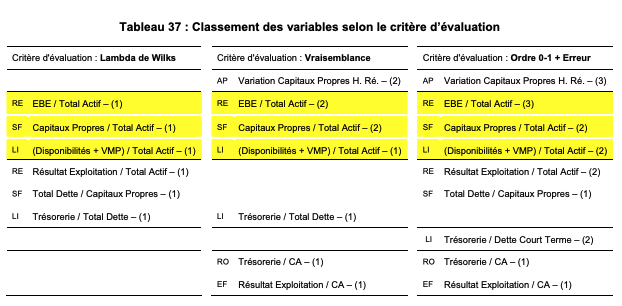
Objectif de la semaine du 28 mars : à la fin de la semaine, ce serait top d’avoir quasi fini tout ce qui se fait dans un notebook.

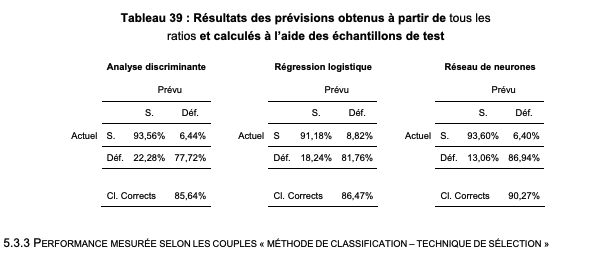
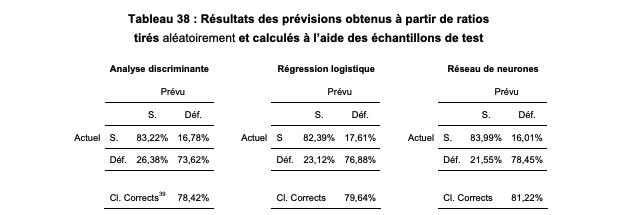
* Faire une baseline
* Améliorer le modèle
  + Feature Engeneering
* Travailler sur les données
  + Sélection des p
* Tracer les expériences (a minima dans un tableau, au mieux dans ML Flow)
* Ne conserver que les variables correspondant aux ratios, auxquelles il faut rajouter :
  + Le code APE
  + L’effectif de l’entreprise
  + L’âge de l’entreprise
  + Des montants en valeur (une notion de taille à affiner pour prédire le niveau de risque, peut-être prévoir des classes ?)
    - Total du bilan
    - Total des KP
    - Total du CA
    - Total du RCAI

Analyse des QQ plots et histplot :

|  |  |  |
| --- | --- | --- |
| Crédit client | Asymétrique à droite + il semble y avoir des outliers (tout est autour de 0 mais le | * Supprimer les outliers en + et relancer pour voir. |
| Crédit fournisseur | Grosses queues de distribution, surtout à droite | * Supprimer les outliers en + (5000 déjà ?) et en - |
| Rotation stocks | Asymétrique à droite + qq outliers |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

* Supprimer les variables dont on est sûr qu’elles ne vont pas servir (après calcul des ratios)
* Etudier la distribution des variables qui restent (distribution normale ? outliers ?)
* Etudier la relation qu’elles entretiennent avec la cible (scatterplot ou boxplot, + tests de différence de moyenne => test t, test F, test non paramétrique U de Mann-Whitney, test de normalité de la distribution ou qqplot)
* Critères de sélection des variables :
  + Critère Lambda de Wilks (analyse discriminante)
  + Critère de vraisemblance
* Tenter de





|  |  |  |
| --- | --- | --- |
| Variable | Descriptif | Tests |
| CA | 2302 entreprises avec un CA null  Les ratios correspondants sont soit en Nan soit en infinite.  Est-ce pertinent de garder ces entreprises ?   * A tester : est-ce que la distribution entre des entreprises ayant un CA null et celles n’en ayant pas est-elle significativement différente ? |  |
| Date de création | * A passer en nombre d’années depuis la création |  |
| Tranche d’effectif | * Voir s’il y a un impact de ces valeurs sur la target (boxplots ?) * Trouver un moyen d’imputer les valeurs manquantes (12%) * Voir s’il est pertinent de faire des regroupements de catégories |  |
| Code APE | * Regrouper dans de plus petites catégories ? * Identifier peut-être quelques secteurs particulièrement touchés (restauration, autres ?) |  |
| Catégorie juridique | * Soit supprimer les données, soit imputer les valeurs manquantes de la catégorie de niveau 2 |  |
| Code postal  Ville | * Trop de modalités. Mais peut être intéressant (taille de la ville, nombre d’habitants dans un certain périmètre autour ?) * Pourquoi pas dans un premier temps département, ou région ? * Table de correspondance INSEE : * A grouper avec table population au niveau le plus fin (commune) pour analyse | Pour l’instant supprimer (ou ne prendre que les départements, puis aviser)  Ca sera la perspective d’une amélioration future (analyse géographique des risques de défaillance : selon les secteurs d’activité, selon e type d’implantation aussi : campagne, urbain, périurbain…) |
|  |  |  |
| Variables à supprimer | (nombre de modalités trop important ou pas assez, ne permettant pas d'apporter ) : ["Dénomination", "Procédures collectives (type)", "Procédures collectives (date)", "Ville", "Notation NOTA-PME 2016"] |  |