Data Analysis Checklist

0) Notes préalables : automatisez, automatisez, automatisez. Notez toutes vos réflexions et vos hypothèses. Lister toutes les hypothèses que vous avez effectué dans un document.

1) Définir la problématique, le contexte, la big picture. ~~Comment votre étude / modèle de ML va être utilisé ? Quelles sont les solutions actuelles ? Si problème de ML, supervisé ou non supervisé ?~~

2) ~~Définir une métrique cible. Comment cette métrique cible est liée avec un objectif business ? Quelle serait la performance minimale pour que votre solution soit utilisée ?~~

3) ~~Existe t il des problèmes comparables ? Comment puis je m'inspirer de ces problèmes ?~~

4) ~~Dans le cas de ML : comment ferais je si je devais faire une approche manuelle ?~~

5) Lister la donnée requise

6) Trouver et documenter où est la donnée : nature, source, donnée, accès légal.

7) Splitter le dataset en 2 parties : 80/20. 80% sur laquelle vous faites votre analyse exploratoire, 20% sur laquelle vous faites vos confirmations de tests statistiques. ~~Splitter le dataset en 3 parties : une partie pour l'exploration et l'entraînement, une partie pour la validation du meilleur modèle de ML supervisé le cas échéant, et une partie pour la confirmation (à cause du data dredging) et le test pour l'évaluation (NE JAMAIS REGARDER CETTE PARTIE AVANT L'EVALUATION) : 80/10/10 par exemple. Le mieux restant de la cross-validation.~~

8) Ecrire le codebook de la donnée :

    1) Information sur la nature, l'origine du dataset, la manière dont la donnée a été collectée, etc.

    2) Informations sur les variables :
        1) Nom de la variable
        2) Type de variable : numérique, catégoriel, (ordinal)
        3) Segmenter les variables en grandes catégories (ex : spatiales, temporelles, etc.)
        4) Notre attente a priori sur l'influence de la variable par rapport à la variable cible (High, Medium, Low)
        5) % of missing values
        6) Distribution : normal shaped, skewed, long tailed ?
        7) Outliers ?
        8) Transformation mathématiques potentiellement intéressantes
        9) Autres commentaires

    3) Rajouter au fur et à mesure l'information sur des variables rajoutées

9) Préparer la donnée :

    1) Fix or remove outliers (optionnel)
    2) Missing values :
        1) Très peu de valeurs manquantes (moins de 1%) : on peut dropper les lignes
        2) Entre 1 et 10% : Fill in missing values (avec 0, moyenne ou médiane)
        3) Plus de 10 % drop the column
    3) Feature selection : drop the attributes that provide no useful info for the task
    4) Rajout de variables potentiellement intéressantes
    5) Transformation mathématiques potentiellement intéressantes
    6) Si beaucoup de data, potentiellement faire un sampling aléatoire pour réduire la taille du dataset

10) Explorer la donnée :
    1) Faire le plus de graphes le plus rapidement possible. On se fiche pour le moment que les graphes soient jolis.
    2) Sortir des statistiques avec des intervalles de confiance si peu de datas.
    3) Noter au fur et à mesure les leçons apprises
    4) Graphes univariés (une seule variable) : Histogrammes / Density Plot, Box Plot, Bar Plot
    5) Explorer les relation entre 2 variables :
    6) Graphes multivariés : Scatter Matrix, Plot, Heatmap, Table de contingence, Box Plot "stackés"
    7) Matrices de corrélations
    8) ~~Matrices de chi square pour les variables catégorielles~~
    9) Si potentiellement relations linéaires, développer avec régression linéaire
    10) Si potentiellement relation non linéaire, tester des transfo mathématiques pour linéariser le problème
    11) Différence de 2 populations : t-test
    12) Différence de N populations : anova
    13) ~~Indépendance entre variables catégorielles : chi square~~
    14) Conclure sur les potentielles relations entre variables

11) Confirmer les hypothèses en testant la p-value sur le dataset de confirmation (cf 8)

12) ~~ML Supervisé : Short-list des modèles les plus prometteurs~~
    1) ~~S'il y a bcp de données, vous voudrez peut-être échantillonner des ensembles d'apprentissage plus petits afin de pouvoir entraîner de nombreux modèles différents dans un délai raisonnable (sachez que cela pénalise les modèles complexes tels que les grands réseaux de neurones ou les forêts aléatoires). Encore une fois, essayez d'automatiser ces étapes autant que possible~~
    2) ~~Entraînez de nombreux modèles rapides et sales de différentes catégories (par exemple, linéaire, Bayes naïf, SVM, forêts aléatoires, réseau neuronal, etc.) en utilisant des paramètres standard.~~
    3) ~~Mesurez et comparez leurs performances. Pour chaque modèle, utilisez la validation croisée N fois et calculez la moyenne et l'écart type de la mesure de performance sur les N plis.~~
    4) ~~Analysez les variables les plus significatives pour chaque algorithme~~

5) ~~Analysez les types d'erreurs commises par les modèles. Quelles données un humain aurait-il utilisées pour éviter ces erreurs?~~
6) ~~Faites un tour rapide de sélection des fonctionnalités et d'ingénierie~~
7) ~~Ayez une ou deux itérations rapides supplémentaires des cinq étapes précédentes~~
8) ~~Faites une liste restreinte des trois à cinq modèles les plus prometteurs, en préférant les modèles qui font différents types d'erreurs.~~

13) ~~Fine tuning. Remarques: Vous souhaiterez utiliser autant de données que possible pour cette étape, en particulier lorsque vous vous rapprochez de la fin de la mise au point. Comme toujours, automatisez ce que vous pouvez.~~
    1) ~~Affinez les hyperparamètres à l'aide de la validation croisée.~~
    2) ~~Traitez vos choix de transformation de données comme des hyperparamètres, en particulier lorsque vous n'en êtes pas sûr (par exemple, dois-je remplacer les valeurs manquantes par zéro ou par la valeur médiane? Ou simplement supprimer les lignes?).~~
    3) ~~À moins qu'il n'y ait très peu de valeurs d'hyperparamètres à explorer, préférez la recherche aléatoire à la recherche par grille. Si la formation est très longue, vous pouvez préférer une approche d'optimisation bayésienne (par exemple, en utilisant des priors de processus gaussiens, comme décrit par Jasper Snoek, Hugo Larochelle et Ryan Adams). 1~~
    4) ~~Essayez les méthodes Ensemble. La combinaison de vos meilleurs modèles fonctionnera souvent mieux que de les exécuter individuellement.~~
    5) ~~Une fois que vous êtes sûr de votre modèle final, mesurez ses performances sur l'ensemble de test pour estimer l'erreur de généralisation~~
    6) ~~AVERTISSEMENT Ne modifiez pas votre modèle après avoir mesuré l'erreur de généralisation: vous commenceriez simplement à surappuyer l'ensemble de test.~~

14) Préparer la présentation :

    1) Graphiques :
        1) Est-ce que chaque figure communique UNE info importante ?
        2) Est-ce que les axes sont bien présents, la légende, l'échelle ?
        3) Est-ce que les figures sont lisibles ?

    2) Présentation :
        1) Avoir un pitch en une ligne ou moins de 20 secondes du sujet
        2) Décrire la problématique
        3) Décrire la donnée, le dataset, que représente une ligne, etc.
        4) Storyteller comme si vous racontiez votre enquête.
        5) Communiquez sur chaque slide avec UNE idée principale. Ex : "the median income is the number-one predictor of housing prices"
        6) Conclusion, Going Further