3.1 L'ingénierie des features est importante ::

Le choix des procédés d'extraction des features est important et influence sur le résultat obtenu, il y a souvent des features cachées (latentes) dans les données qui peuvent améliorer les performances du modèle, cependant avoir trop de features peut réduire la précision du modèle à cause de la redondance et de l'inutilité de certaines d'entre-elles.

Pour les données textuelles et les images, les features peuvent être extraites automatiquement par des modèles de Deep-learning.

3.2 Les modèles simples sont compétitifs :

Un modèle simple est facile à entrainer, à modifier et à comprendre, comme il peut être adapté à différentes situations, cependant, il demande plus de traitement sur les données elles-mêmes (nettoyage de données, valeurs manquantes, etc.).

La simplicité peut-être vue non seulement en fonction de la complexité du modèle, mais aussi en tenant compte du prétraitement des données : parfois une simple régression linéaire régularisée peut faire la différence.

3.3 La combinaison de modèles est une stratégie gagante :

L'apport combiné de nombreux modèles dont la corrélation est faible améliore certainement la précision prédictive, est une stratégie gagnante quoique conseillée en dernier ressort.

3.4 Overfitting peut être un problème :

Les paramètres et hyper-paramètres d'un modèle doivent être appris en cross validation sur les données d'apprentissage afin d'avoir un meilleur score.

Une validation croisée correcte a montré son efficacité pour résoudre l'un des problèmes majeurs (le sur-apprentissage).

3.5 Bien prédire est important:

Une des façons d'aborder le problème de prédiction est soit de collecter le maximum d'informations correctes dérivées de l'objet à prédire ou sinon d'entrainer les modèles statistiques en choisissant le bon critère d'évaluation, la fonction de perte appropriée (optimiser la métrique correcte).