Session de formation - Machine learning

Partie 2 – Dasek Joiakim

1. Gestion des Valeurs Manquantes

- **Approche des valeurs manquantes** : Deux points de vue ont été discutés concernant la gestion des valeurs manquantes : remplacer les valeurs nulles par la moyenne de la colonne ou les gérer comme des "NaN" (Not a Number).

- **Risques du remplacement par la moyenne** : Remplacer les valeurs manquantes par la moyenne peut induire en erreur une analyse puisqu'il pourrait introduire des corrélations factices entre des données qui, en réalité, ne le sont pas.

- **Statistiques à utiliser** : Les approches doivent inclure des statistiques qui prennent en compte les "NaN" au lieu de les remplacer. En cas d'absence de valeurs, ignorer complètement ces points peut parfois être la meilleure option.

2. Modèles et Validation

- **Importance du dataset** : L'exhaustivité et la qualité des données sont cruciales. Utiliser des données incomplètes ou erronées peut compromettre la fiabilité des résultats.

- **Validation croisée** : Analyser les relations entre les données à l'aide de techniques de validation croisée, comme le R^2 (coefficient de détermination), pour évaluer l'ajustement du modèle.

- **Sur-optimisation** : L'utilisation excessive de la moyenne ou des solutions simplistes peut conduire à des modèles trop ajustés aux données, au détriment de la généralisation.

3. Traitement et Manipulation des Données

- **Sélection des caractéristiques** : La sélection des variables influentes pour l'oxygène a été abordée. Par exemple, l'importance du pH, de la température, etc., sur l'oxygène dissous a été discutée.

- **Clustering et Analyse Multidimensionnelle** : Des techniques de clustering peuvent être appliquées pour identifier des modèles dans de grands jeux de données, tandis que des méthodes de réduction de dimensionnalité (ex. UMAP) peuvent aider dans la visualisation.

4. Modèles de Machine Learning

- **Random Forest et Régression** : Le modèle Random Forest a été mentionné comme imparfait face aux valeurs manquantes. La nécessité de s'assurer que tous les modèles sont formés avec des jeux de données propres a également été soulignée.

- **Comportement des modèles** : Les différents modèles peuvent avoir des comportements variés selon la structure des données et la présence de valeurs manquantes ou de bruit.

5. Approches de la Data Science

- **Interprétation des résultats** : Il est essentiel d'interpréter les résultats au sein du contexte commercial pour éviter de fausses conclusions basées sur des corrélations triviales.

- **Analyse Causale vs Corrélationnelle** : Une distinction claire entre l'identification de tendances causales et l'observation de corrélations est primordiale. Les données doivent être analysées avec rigueur afin de garantir leur pertinence.

6. Données et Dynamique des Modèles

- **Importance de la Compréhension Business** : La compréhension des enjeux d'affaires ainsi que des meilleures pratiques en matière de data science et machine learning est cruciale pour orienter les décisions basées sur les données.

- **Fin de la session** : La discussion s'est achevée sur l'importance d'adopter une approche itérative et adaptative face à l'évolution des techniques d'analyse de données et des circonstances des modèles utilisés.

Conclusion

La gestion des données, la rigueur scientifique et l'adaptation aux besoins spécifiques des utilisateurs sont des éléments clés pour réussir dans le domaine de la data science. Les discussions autour des meilleures pratiques soulignent l'importance d'analyser en profondeur les implications de chaque méthode employée dans les modèles analytiques.