Sommaire

A propos

Ce document présente les différentes étapes par lesquelles vous devriez impérativement passer quel que soit votre projet.

En gros, le projet est un TP plus détaillé de ce qu'on a vu dans le use case du TP 3.

Voyez cette notice comme un fil conducteur si vous ne savez pas par où commencer et/ou si vous êtes perdus en cours de route. A la fin de cette note vous trouverez aussi quelques informations sur ce qui est attendu.

Charger et explorer les données

La première étape consiste à récupérer la base de données, la mettre sous forme de Dataframe avec pandas et d'explorer le contenu. Vous pouvez regarder le TP2 et le TP3 pour la lecture, l'affichage et la manipulation des dataframes.

Cette étape est extrêmement importante, assurez-vous avant d'aller plus loin de savoir correctement répondre aux questions suivantes :

- Quantité de données :
    - Quel est le type de données par colonne ?

- Quelle est la taille de la base de données (en nombre de lignes et de colonnes, etc.) ?
- Quelles sont les colonnes exploitables/importantes pour votre étude statistique ?
- etc.

Nettoyage et présentation de données

Pour la compréhension et l'exploration des données on regarde la qualité des données : * Les données comprennent-elles des caractéristiques pertinentes pour la problématique ? * Quels sont les types de données présents (symbolique, numérique, etc.) ? * Avez-vous calculé des statistiques de la base pour les colonnes clés ? * En quoi cela va-t-il permettre d'éclaircir la problématique ? * Est-il possible de ne garder que les colonnes pertinentes ?

Voici d'autres questions que vous pouvez vous poser :

- Quels sont les colonnes qui semblent sans intérêt et peuvent être exclus ?
- Le nombre de données (échantillion) permet-il de tirer des conclusions pouvant être généralisées ou d'effectuer ?
- Avez-vous envisagé le mode de traitement des valeurs manquantes dans chacune de vos sources de données ?
- Avez-vous exploré les écarts afin de déterminer s'il existe des valeurs aberrantes ou des phénomènes à analyser plus en profondeur ?
- Avez-vous envisagé d'exclure les données sans impact ?
- Avez-vous utilisé des graphiques exploratoires pour clarifier les attributs-clés ?
- Savez-vous quels sont les attributs à fusionner ?

Voilà un exemples de traitements si besoin :

Problème posé par les données	**Solution possible**
Données manquantes	Exclure les lignes ou les caractéristiques, ou insérez une valeur estimée dans les blancs.
Erreurs dans les données	Procédez de manière logique pour découvrir manuellement les erreurs et les corriger, ou exclure les caractéristiques.
Codage des incohérences	Décidez d'une méthode de codage unique, puis convertissez et remplacez les valeurs.
Métadonnées erronées ou manquantes	Examinez manuellement les champs suspects et recherchez la signification correcte.

Pour cette partie, on peut utiliser les méthode pandas pour visualiser des histogrammes, des boîtes à moustaches, des camemberts, etc. On peut aussi facilement calculer les moyennes, les médianes, les variances et bien sûr les quantiles.

Pour faire simples, on considérera les données comme aberrantes si elles sont inférieures au quantile 5% ou supérieures au quantile 95%.

Modèles de régression simple

Vous avez un exemple détaillé dans le TP3.

Il s'agit d'étudier la corrélation entre les colonnes choisies de la Dataframe. On peut créer une nouvelle dataframe qui ne contient que les colonnes choisies, si besoin.

Il existent plusieurs librairies qui permettent de faire une régression : sklearn (vue en TP), statsmodels (https://www.statsmodels.org/stable/index.html) et bien d'autres. L'objectif est : * choisir les variables explicatives et la variable à expliquer ; * choisir un modèle de régression et de bien le comprendre ; * afficher les résultats et les interpréter.

Au final qu'est-ce qui est attendu ?

1. Ouvrir, afficher, comprendre ses données.
2. Explorer les données à l'aide de graphiques.
3. Nettoyer et préparer les données.
4. Repérer les colonnes importantes.
5. Choisir les variables explicatives et la variable à expliquer : faire une régression et commenter les résultats.
6. BONUS : comparer plusieurs modèles en sachant identifier les avantages et les inconvénients de chacun.

Quelques conseils pour une bonne maîtrise

- Ne brûlez pas les étapes : assurez-vous d'avoir compris toutes les parties de tous les TPs avant de commencer, prenez le temps de comprendre vos données.
- Assurez vous de comprendre la régression .
- Commencez simple, par exemple si vos données possèdent plusieurs attributs, essayer avec 1-2 au début et avec une régression, ensuite, ajoutez des attributs.
- Mieux vaut une solution simple qui fonctionne qu'une méthode complexe que vous ne maîtrisez pas.
- Venez aux avec des questions.