La préparation et la qualification des données

Brahiman & Xavier

1. Définitions

LA PREPARATION DES DONNEES

La préparation des données est l'une des phases les plus importantes dans le cycle de vie d'un projet data science. Les données brutes étant souvent bruyantes, peu fiables et incomplètes, leur utilisation pour la modélisation peut générer des résultats trompeurs. Le TDSP (Team Data Science Process) consistant en l'exploration initiale d'un jeu de données utilisé permet de découvrir et planifier le traitement préliminaire requis conduisant au nettoyage et à la transformation des données avant leur traitement. Plusieurs solutions existent.

LA QUALIFICATION DES DONNEES

C'est une étape qui consiste en la collecte et au stockage de données de bonne qualité dans la base de données

2. Les techniques de préparation des données

La préparation des données dans un projet de data science est essentiels.

Les données présents dans les jeu de données ne sont pas toujours bien structurés, ni complètes, ni bien formatées.

Dans ce cas, il est donc indispensable de passer par l'étape de nettoyage et de transformation des données.

Les techniques de préparation sont nombreuses, à tel point qu'on ne peut pas être sûr d'utiliser la bonne méthode.

Ces méthodes varient suivant la nature des projet ainsi que des données utilisées.

2.1 Traitement des valeurs manquantes

Comme nous l'avons remarqué précédemment, il est possible que le jeu de données que nous traitons soit parsemer de valeurs proches de zéro. Ces valeurs proches de zéros peuvent être cohérentes suivant l'ordre de grandeurs des toutes les valeurs du jeu de données. Imaginons que ces valeurs soit incohérentes et que nous devons effectuer un trie de celles ci

Ce phénomène arrive très fréquemment en pratique et peut s'expliquer par diverses raisons, allant de la panne de capteurs à la mauvaise remontée des mesures (et même parfois pour des raisons inexpliquées).

Cependant, lorsque nous étudions des séries temporelles, laisser des mesures aberrantes ou manquantes dans le jeu d'entraînement peut avoir des répercussions néfastes sur les résultats obtenus par les prévisions qui s'en suivent.

C'est pourquoi il est important d'avoir la possibilité de remplacer ces valeurs par des valeurs plus vraisemblables. On appelle cela **l'imputation**.

2.2 L’imputation sur le démonstrateur

Il est possibles d’imputer les valeurs de plusieurs façons, en voici trois différentes :

- **Remplacer les valeurs manquantes par la moyenne sur le jeu d’entraînement**. Cette méthode permet d’éviter d’induire en erreur les modèles de prédiction en supprimant les valeurs éloignées du jeu de données. Elle convient uniquement lorsque la série ne contient qu’un faible taux de valeurs manquantes.
- **Ajustement saisonnal et d’interpolation linéaire**. Son principe est de remplacer les valeurs manquantes par une valeur calculée par la désaisonnalisation de la série, suivie d’une interpolation linéaire et d’une “resaisonnalisation” de cette même série. L’objectif de cette méthode plus avancée est de prendre en compte la saisonnalité et la tendance de la série que nous souhaitons imputer.
- La dernière méthode est une méthode issue d’un algorithme de Machine Learning : le **Gradient Boosting**. Ce modèle sera aussi proposé en tant que modèle pour la prévision et il possède la particularité de s’appuyer sur les variables externes de la série et non sur ses comportements temporels comme c’était le cas pour les méthodes précédentes.

2.3 Voici trois modèles différents pour la phase de modélisation.

- **Holt-Winters**. C'est un modèle statistique basé sur la **décomposition additive** d'une série temporelle (tendance, saisonnalité, et résidus). Ce modèle est simple ainsi que son fonctionnement. Il peut être très performant dans certaines situations. Sa principale limitation est qu'il ne possède pas la capacité d'inclure des variables exogènes à la modélisation, comme les données météorologiques.
- Un modèle dit de **Prophet**. Il a été développé et mis en Open Source par **Facebook** qui l'utilise notamment en interne pour ses routines quotidiennes de prédiction. Tout comme Holt-Winters, il permet la **décomposition additive** d'une série temporelle, mais en incorporant des cycles saisonniers de façon plus avancée (cycles annuels, hebdomadaires, quotidiens, mais aussi des effets vacances par exemple), ainsi qu'en offrant la possibilité de prendre en compte des **variables exogènes** dans la modélisation.
- Pour finir, le dernier modèle est un modèle de Machine Learning, le **modèle de Gradient Boosting**. Très différent des modèles statistiques précédents dans sa composition, le Gradient Boosting est un **algorithme d'apprentissage supervisé de type ensembliste** dont le principe est de combiner les résultats d'un ensemble de modèles plus simples et plus faibles afin de fournir une prédiction globale très fiable. Contrairement aux modèles précédents, ce modèle n'est pas intrinsèquement conçu pour extrapoler les valeurs d'une série temporelle. Toutefois, en lui fournissant des variables porteuses d'information sur la tendance ou la saisonnalité de la série par exemple, il est capable de capturer des corrélations avancées permettant de prédire les valeurs de la série.

2.4. Les bases de la prévision de séries temporelles

Lorsque l'on souhaite effectuer la prévision d'une série temporelle, les principales étapes à suivre sont :

- La **séparation de la série en deux** : un jeu d'entraînement sur lequel nous entraînons notre modèle à reconnaître les différents comportements de la série, et un jeu de test sur lequel nous effectuons des mesures de la cohérence de nos prévisions. La date charnière sera appelée la « date de cutoff ». Il est très simple de la faire varier sur notre démonstrateur.
- L'**analyse du jeu d'entraînement** afin d'identifier ses différents comportements notables : la série possède-t-elle des motifs saisonniers ? Une tendance ? Est-ce que des valeurs aberrantes sont présentes et risquent de fausser les prévisions ?
- Dans le cas où le jeu d'entraînement comporte des valeurs manquantes ou aberrantes, il est nécessaire d'effectuer un traitement préalable sur les données.
- Vient ensuite la **phase de modélisation**, lors de laquelle nous comparons un ensemble de modèles selon différents critères de performance. Une fois le modèle le plus adéquat élu, nous effectuons des tests sur sa robustesse en faisant varier la date de cutoff.
- Enfin, une fois les modélisations effectuées, nous étudions l'influence des différentes variables sur le comportement du modèle pour identifier les plus influentes. Nous vérifions aussi l'incertitude liée aux prévisions en analysant l'**intervalle de confiance à 95%** associé à chaque mesure.

3. La qualification des données

La qualification de données est une opération visant à améliorer la qualité des données recueillies (fichiers clients, prospects, ...) pour les rendre authentiques, complètes et fiables. C'est une technique utilisée dans la prospection, notamment BtoB, et dont l'objectif est d'améliorer la qualité des bases de données afin d'optimiser l'efficacité des actions en découlant (futures campagnes de prospection, ...)

Elle se fait en 4 étapes principales:

- Récupérer des données de qualité dès la récolte
- Enrichir la base de données
- Trier et segmenter les data
- Conserver une base de données à jour

3.1 Récupérer des données de qualité dès la récolte

La meilleur façon d'obtenir des données qualifiées dès la collecte est de recueillir ces données auprès des entreprises (B2B) dont le SIRET seul fourni beaucoup d'informations comme la raison sociale, l'adresse, le statut juridique, ….

Quant aux données des particuliers (B2C), bien évidemment, aucun numéro ne permet de tout connaître d'une personne. Pour qualifier en aval sa base de données, il est donc important en amont de recueillir au moins deux informations : le nom + un critère géographique, qui permet ensuite de se renseigner sur la bonne personne.