Data Scientist - Test Technique Préliminaire

Ce test vise à évaluer des compétences simples en Python et en machine learning. Le rendu doit se faire sous format PDF pour la partie use case et sous forme de fichier python ou jupyter notebook pour la première partie.

I. Test Technique

En tant que Data Scientist missionné par l'entreprise Aidyl (fictive), on vous donne accès à un fichier train.csv contenant des données d'employés de longue date, qui ont passé une évaluation RH de satisfaction. Un autre fichier test.csv vous est également remis. Celui-ci contient les données des salariés les plus récents, qui n'ont pas encore passé d'évaluation de satisfaction. Votre objectif est de développer un modèle permettant d'inférer la satisfaction de ces derniers.

Voici une brève description des données :

- EmployeeNumber : Numéro unique identifiant chaque salarié.
- Satisfaction : Variable binaire encodant le bilan de l'enquête satisfaction pour ce salarié. 1 veut dire que le salarié semble motivé et satisfait de ses conditions de travail. 0 veut dire qu'il est plutôt mécontent. Cette variable est présente uniquement dans les données train.csv et c'est elle que nous souhaitons inférer à partir des autres variables contenues dans test.csv.
- Age : Age actuel du salarié.
- BusinessTravel : Fréquence des voyages d'affaires.
- Department : Département dans lequel travaille le salarié.
- DistanceFromHome : Distance en kms séparant la résidence du lieu de travail.
- Education : Nombre d'années d'études post-bac.
- EducationField : Domaine d'études.
- Gender : Sexe du salarié.
- JobInvolvement : Entier allant de 1 à 4 encodant le niveau d'engagement du salarié au travail.
- JobLevel : Entier entre 1 et 5 encodant sa position hiérarchique.
- JobRole : Poste du salarié.
- MonthlyIncome : Salaire mensuel.
- NumCompaniesWorked : Nombre de sociétés dans lesquelles le salarié a travaillé au cours de sa carrière.
- OverTime : Si le salarié a déjà fait des heures supplémentaires ('Yes') ou pas ('No')
- PerformanceRating : Entier entre 1 et 4 encodant `a quel point le salarié est performant au travail.
- StandardHours : Nombre d'heures de travail réglementaires par semaine.
- StockOptionLevel : Entier entre 0 et 3 encodant la quantité de parts de la société détenus.

- TrainingTimesLastYear : Nombre de formations suivies au cours de l'année précedente.
- WorkLifeBalance : Entier entre 1 et 4 encodant l'équilibre entre vie personnelle et travail.
- YearsAtCompany : Nombre d'années travaillées au sein de la société.
- YearsInCurrentRole : Nombre d'années au poste actuel.
- YearsSinceLastPromotion : Nombre d'années depuis la dernière promotion.
- YearsWithCurrManager : Nombre d'années avec le manager actuel.

I.1 Statistiques Descriptives

L'objectif de cette partie est de comprendre et interpréter les informations contenues dans le jeu de données de façon qualitative.

1. Est-ce que certaines variables vous semblent redondantes (contiennent la même information) ? Si oui, lesquelles ? Quel danger cela peut présenter pour la suite de l'analyse ?
2. Quels facteurs semblent le plus influencer la satisfaction des salariés ?
3. Vérifiez visuellement que les distributions des variables communes aux deux fichiers de données sont similaires. Expliquez en quelques mots pourquoi cela est important. Proposer une méthode de validation numérique de cette assertion.

I.2 Apprentissage

L'objectif de cette partie est de prédire la satisfaction des salariés du fichier test.csv.

1. Quelle métrique choisiriez-vous pour évaluer vos prédictions?
2. Entraînez un modèle de régression logistique sur les données de train.csv et prédire la satisfaction des salariés de test.csv
3. Proposez et entraînez un ou plusieurs autres modèles statistiques pour prédire la satisfaction des employés. Estimez la qualité de vos prédictions et comparez vos estimateurs entre eux. Générez un fichier .csv contenant un tableau avec vos meilleures prédictions.

4. Question bonus : Imaginez maintenant que le fichier train.csv contient 10 000 fois plus d'observations et que les résultats de votre analyse sont attendus pour demain. Expliquez en quelques lignes (15 max.) ce que vous changeriez dans votre approche ?

II. Use Case (1 page)

Vous disposez d'une base de données de grande dimension contenant les noms et des informations sur des marchands.

id	Name	Info
1	Uber	VtC
2	Mc Donalds	restaurant
3	Café st Jean	Bar

On vous demande de mettre en place un algorithme "capable de passer à l'échelle" qui prend en entrée une chaîne de caractère et qui donne en sortie un ensemble de marchands susceptibles de correspondre à cette chaîne. Cette correspondance doit-être quantifiée par un score de matching.

Exemple:

- Entrée: "Uber xxx" -> Sortie : {"id":1, "name":"Uber", "score": 0.99}

Quelles stratégies mettez-vous en place ? Quelles questions cela pose-t-il ? On suppose que vous pouvez changer le stockage, les technologies et que vous pouvez modifier les entrées de la base de marchands.