Robin Sallé 18/11/2024

Zhang Chuqi

Data Science (COVID-19)

Outil(s) IA utilisés : ChatGPT4, Copilot, Gemini

**Section 1 : Introduction et structure du projet**

* 1. Objectifs du projet :

Créer un modèle de classification de clients pour une entreprise automobile afin de l’aider à pénétrer dans un nouveau marché. Le projet consistera à analyser la forme et le fond des données des clients, à transformer ces données afin de les optimiser et à développer un modèle de prédictions pour ensuite le tester.

Nous espérons pouvoir créer le modèle de prédictions le plus optimal possible avec une précision maximale grâce à l’aide des IA.

* 1. Structure du projet :
  2. Lien du Colab : <https://colab.research.google.com/github/robin-salle-vinci/projet_IA/blob/main/BIN3100_Projet_Data_Science.ipynb>

**Section 2 : Apports de l’IA à chaque étape du projet**

**Section 3 : Conclusion et Bilan**

**Annexe : Journal des Interactions avec l’IA**

|  |  |  |
| --- | --- | --- |
| **Date** | **Conversation avec**  **l’IA** | **Résumé de la**  **Réflexion Personnelle et Décision Prise** |
| 18/11/2024 | <https://chatgpt.com/share/673b2009-5128-800e-8af9-5d4672640097> | Il m’a donné beaucoup d’étapes superflues comme la création des branches.  J’ai donc juste suivi l’étape de création du répertoire Github et l’étape de l’automatisation de la collaboration entre Colab et Github. |
| 22/11/2024 | <https://chatgpt.com/share/6740a97d-e448-800e-b3db-c3d0fe58dc1e> | Il m’a donné pas mal de modèle intéressant pour le projet mais je les ai filtré en me référençant aux cheat sheet de SKLearn |
| 22/11/2024 | <https://chatgpt.com/share/6740bbe1-c968-800e-9038-0dd9e8f0524f> | Il m’a conseillé quels modèles de machine learning avaient besoin d’une normalisation de données avec StandardScaler() et il était en adéquation avec ce que Monsieur Saint-Cirgue conseillait. |
| 23/11/2024 | <https://chatgpt.com/share/6741b804-a3c0-800e-875c-3190fb2d3ff1> | Questions sur des concepts de machine learning afin de mieux comprendre le projet. |
| 23/11/2024 | <https://chatgpt.com/share/6741bb9e-54d0-800e-81d5-87bde51d344c> | Il m’a donné les raisonnements pour trouver les meilleurs paramètres d’une learning curve que j’ai mis en application. |
| 25/11/2024 | ton rôle : expert en machine learning avec SKLearn.  énoncé:Une entreprise automobile envisage de pénétrer de nouveaux marchés avec ses produits existants (P1, P2, P3, P4 et P5). Après une étude de marché intensive, elle en a déduit que le comportement du nouveau marché est similaire à celui de son marché existant. Sur le marché existant, l'équipe de vente a classé tous les clients en quatre segments (A, B, C, D). Elle a ensuite mené des actions de sensibilisation et de communication segmentées pour chaque segment de clients. Cette stratégie a exceptionnellement bien fonctionné. Elle prévoit d'utiliser la même stratégie pour les nouveaux marchés et a identifié 1000 nouveaux clients potentiels. Il vous est demandé d'aider le directeur à prévoir le bon groupe de nouveaux clients. Ces données se trouve dans le fichier Customer\_Segmentation.csv se trouvant sur moodle Nous vous demandons de fournir un modèle permettant de faire des prédictions sur le segment dans lequel classer chaque nouveau client.  exemple des données(7000 lignes au total): ID,Gender,Ever\_Married,Age,Graduated,Profession,Work\_Experience,Spending\_Score,Family\_Size,Var\_1,Segmentation 460108,Male,Yes,42,Yes,Artist,1.0,Average,3.0,Cat\_2,C 459869,Female,No,23,No,Healthcare,1.0,Low,3.0,Cat\_6,D 460192,Male,Yes,41,Yes,,0.0,Low,2.0,Cat\_7,D 463582,Female,No,18,No,Healthcare,1.0,Low,3.0,Cat\_6,D 467114,Male,Yes,59,Yes,Marketing,0.0,Average,3.0,Cat\_6,B 461775,Female,Yes,83,No,Lawyer,0.0,High,2.0,Cat\_6,B 463160,Female,No,48,No,Entertainment,,Low,1.0,Cat\_6,B  ma question: Selon toi, quel modèle de SKLearn sera le plus performant et optimal avec cet énoncé et ce genre de données ?  Étant donné la nature des données et l'objectif :   * **Gradient Boosting Classifier (LightGBM ou XGBoost)** sera probablement le plus performant. Ces modèles gèrent bien les interactions non linéaires et sont robustes pour ce type de dataset. * Si vous recherchez un modèle plus simple à interpréter, optez pour **Random Forest**. * Une analyse comparative (GridSearchCV) entre ces deux modèles pourrait affiner la décision finale. | Je lui ai posé cette question pour anticiper le résultat de la fonction d’évaluation entre les différents modèles.  Il m’a donné une réponse qui pourrait selon moi être cohérente. |
| 25/11/2024 | <https://chatgpt.com/share/67446003-0e48-800e-9ae9-fb88ccd5d23d> | Je lui ai demandé les hyper-paramètres les plus important pour le modèle GradientBoostingClassifier qui pourrait être le modèle le plus optimal pour ce projet.  J’ai récupérer ces réponses tout en regardant la doc sklearn pour être sur |