Sallé Robin 18/11/2024

Zhang Chuqi

Data Science project

Outil(s) IA utilisés : ChatGPT-4, Copilot, Gemini (Colab)

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

* 1. Objectifs du projet :

Créer un modèle de classification de clients pour une entreprise automobile afin d’aider l’entreprise pour pénétrer de nouveau marché. Le projet consistera à analyser la forme qui consiste l’identification de la variable cible, dimensions du jeu de données, analyse des valeurs manquantes et des types de variables et l’analyse du fond qui consiste en la visualisation de la distribution de la variable cible, recherche d'informations sur le domaine pour comprendre la signification des variables, visualisation des relations entre les variables prédictives et la variable cible, et identification des valeurs aberrantes 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 accuracy de 100% grâce à l’aide des IAs.

* 1. Structure du projet :

Nous avons développé ce projet en python sur le site Colab de Google qui permet d’offrir un environnement d’exécution du code avec un CPU ainsi qu’un GPU. Pour ce qui est de la conception du modèle, nous utilisons la librairie SKLearn. De plus, nous utilisons matplotlib.pyplot pour visualiser les scores de notre modèle via des graphiques.

* 1. Lien du Colab :

<https://colab.research.google.com/github/robin-salle-vinci/projet_IA/blob/main/BIN3100_Projet_Data_Science.ipynb>

Lien du répertoire Github :

<https://github.com/robin-salle-vinci/projet_IA>

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

* Idéation et Conception :
* Développement :
  + chatGPT-4 nous a permis d’apprendre l’utilisation d’un répertoire distant Github avec un Colab, de choisir notre modèle de machine learning en fonction de la demande du projet, d’avoir des éclaircicements sur des concepts liés aux Data Sciences et de paramétrer nos modèles pour de meilleures performances.
  + Gemini intégré à Colab nous a permis de résoudre 80% de nos messages d’erreurs due à notre code.
  + L’IA nous a aidé à identifier les paramètres adéquats pour imputer notre code
* Validation et Tests :
  + chatGPT-4 nous a permis de développer le code permettant de créer le fichier csv contenant les prédictions de notre modèle et d’implémenter le code permettant de visualiser nos résultats via des graphiques matplotlib.

**Section 3 : Conclusion et Bilan**

L’apport de l’IA pour notre projet a globalement été assez significatif. En effet, en utilisant l’IA de manière beaucoup plus **structurée** et réfléchie (attribution d’un rôle, explication du contexte, …), cela nous a permis de recevoir des réponses plutôt cohérentes en général. La plupart de nos problèmes ont été **résolus** grâce à l’IA que ce soit Gemini qui pouvait directement analyser notre code dans Colab ou encore ChatGPT.

Par exemple, lors de la phase de l’analyse, l’IA nous a aidé à confirmer notre choix d’analyse concernant la visualisation de la variable cible. En ce qui concerne le développement, l’IA s’est avérée précieuse pour le prétraitement des données. Plus précisément, elle nous a guidés dans l’identification des paramètres adéquats pour notre fonction d’imputation. En effet, l’IA a permis d’affiner le filtrage des données pour le traitement des valeurs manquantes, en proposant des seuils et des méthodes d’imputation adaptés à la nature de nos données et aux objectifs de notre modèle. Enfin, l’IA a contribué à la résolution de problèmes techniques lors de la phase de modélisation. Elle a notamment détecté des erreurs subtiles dans le lien entre le prétraitement des données et la construction du modèle, nous permettant ainsi de corriger des incohérences et d’assurer la fluidité du processus.

Cependant, il a été remarqué que ChatGPT avait tendance à s’égarer de la question ou à donner trop d’informations futiles/ **inadaptées**. C’est pourquoi, nous avons **dû** continuellement filtrer ses réponses.

**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 |
| 30/11/2024 | <https://chatgpt.com/share/674af774-fd98-800e-8d06-354b24774eea> | J’avais quelques incompréhensions concernant la precision recall curve et je me demandais si elle était utile pour une classification multi-classes.  ChatGPT m’a répondu que ça pouvait être utile si il y avait une classe désiquilibrée ou d’intêret spécifique.  Ce qui n’est pas le cas seulement moi pour le projet. |
| 30/11/2024 | <https://chatgpt.com/share/674b056b-8474-800e-b8c5-d0c606c4d0f4> | Je lui ai demandé d’implémenter la partie finale de prédictions des données du fichier Customer\_Segmentation\_test.csv afin d’avoir une idée globale de comment procéder + comment créer un fichier csv. |
| 30/11/2024 | <https://chatgpt.com/share/674b4cf8-1f14-800e-90ce-55dde3becb91> | Il m’a aidé à prendre une décision si je devais garder la variable work\_experience qui avait le plus grand taux de valeurs manquantes (10%). |
| 25/11/2024 | https://chatgpt.com/share/674c8ca1-85c8-800b-aa0d-1616459705fd | L’ia m’a aidé à mieux voir plus en détails les codes vu dans les vidéos |
| 29/11/2024 | https://chatgpt.com/share/674c8c2e-5178-800b-be10-9268b52ef059 | L’ia m’a aidé à adapter ma version de code par rapport au codes vu dans les vidéos |
| 01/12/2024 | Moi : voici notre projet Data Science, malheureusement, nos modèles ont un score de 46% d'accuracy (tous les modèles). Dès lors, remarques-tu des problèmes ? ou améliorations à faire sur base de l'analyse faites ?  Gemini : Votre code semble bien structuré et suit les étapes principales d'un projet de Data Science. Cependant, quelques points peuvent être améliorés pour potentiellement augmenter la performance de vos modèles :  **Pre-processing**   * **Imputation des valeurs manquantes**: Supprimer les lignes avec des valeurs manquantes peut entraîner une perte d'informations. D'autres méthodes d'imputation pourraient être plus efficaces (remplacer par la moyenne, la médiane, ou utiliser des algorithmes plus sophistiqués comme KNNImputer). * **Encodage**: Pour les variables ordinales comme 'Spending\_Score', un encodage numérique simple (0, 1, 2) pourrait introduire un ordre artificiel. L'utilisation de One-Hot Encoding pourrait être plus appropriée. * **Feature Engineering**: Il pourrait être intéressant de créer de nouvelles variables à partir de celles existantes. Par exemple, créer des combinaisons de variables ou des termes polynomiaux. * **Feature Scaling**: Vous utilisez StandardScaler, mais d'autres méthodes comme MinMaxScaler pourraient être testées.   **Modélisation**   * **Choix des modèles**: Vous avez testé plusieurs modèles, ce qui est une bonne pratique. Il serait intéressant d'explorer d'autres algorithmes, comme les réseaux de neurones ou les SVM non linéaires. * **Hyperparameter Tuning**: Vous utilisez GridSearchCV, mais il est important d'explorer un espace d'hyperparamètres suffisamment large pour trouver les meilleurs paramètres pour chaque modèle. * **Analyse des erreurs**: L'analyse de la matrice de confusion et des cas mal classés peut donner des indications sur les faiblesses du modèle et guider les améliorations.   **Analyse exploratoire**   * **Analyse de fond**: Il serait intéressant d'approfondir l'analyse des relations entre les variables et la cible. Des visualisations plus avancées pourraient révéler des informations utiles. * **Outliers**: L'analyse des outliers est mentionnée, mais il n'est pas clair si des actions ont été prises pour les gérer.   **Remarques générales**   * **Documentation**: Il est important de documenter davantage les choix effectués à chaque étape, notamment pour le prétraitement et la modélisation. * **Réduction de dimensionnalité**: Si le nombre de variables est important, des techniques de réduction de dimensionnalité (PCA, LDA) pourraient être envisagées.   **Conseils**   * **Itérer**: Le processus de construction d'un modèle de Machine Learning est itératif. Il est important de revenir aux étapes précédentes (analyse exploratoire, prétraitement) en fonction des résultats obtenus lors de la modélisation. * **Expérimenter**: N'hésitez pas à tester différentes approches et à comparer les résultats.   En appliquant ces suggestions et en itérant sur le processus, vous devriez pouvoir améliorer la performance de vos modèles. N'hésitez pas à me poser des questions plus spécifiques sur votre projet si besoin. | Nous avons essayé d’optimiser notre modèle grâce à l’IA Gemini.  Nous avons tester certaines de ses suggestions mais elles se sont avérées infructueuses.  Nous avons seulement réussi à pousser l’accuracy du modèle à 53%. |