Groupe Canam | *Intelligence Artificielle*

Étude de Cas Cameroun

Novembre 2023

CANAM

Objectif du Concours

L'objectif principal est de résoudre une étude de cas en science des données. Les participants disposent d'une fenêtre d'une semaine pour compléter l'étude de cas et soumettre leurs livrables. Ce livrable est constitué des réponses aux questions et le code produit dans un document PDF ou un Jupyter Notebook.

Il s’agit d’un travail individuel qui a pour objectif d’évaluer les compétences suivantes:

- Fondement de la programmation
- Statistiques et mathématiques
- Gestion des données
- Base de données et SQL
- Visualisation des données
- Fondements de l’apprentissage automatique
- Débrouillardise et autonomie
- Capacités de résolution de problèmes
- Compétences en communication / vulgarisation

Étude de Cas

- **Instructions :** Bienvenue au concours CANAM et plus spécifiquement à l'épreuve en Intelligence Artificielle! Ton prochain défi sera de nous démontrer certaines de tes compétences au moyen d'un court cas en science des données. Nous t'avons transmis un jeu de données provenant de la plateforme Kickstarter (voir ficher « ks_dataset.csv »). En utilisant ces données, nous t'invitons à résoudre la problématique ci-dessous. Une fois résolue, renvoie-nous tes réponses aux questions ainsi que le code que tu as produit. Bonne chance!

Lien vers les données (mot de passe: moovai) **CLIQUEZ ICI**

- **Mise en situation :** Certains promoteurs de projet tentent de comprendre comment augmenter le taux de réussite de leurs futures campagnes. Ils ont à leur disposition des données historiques de campagnes annoncées sur la plateforme Kickstarter.

- **Objectif :** Développe en Python une approche ML (supervisée et/ou non supervisée) pour aider les promoteurs de projet à lancer des campagnes à fort potentiel de réussite.

Étude de Cas

- **Question #1 : Préparation des données (20 points)**
  - 1.1. Si tu rencontres des problèmes de qualité des données durant ta manipulation des données de Kickstarter que nous t'avons fourni (indice: nous te confirmons que les données ont été corrompues 😉), comment les as-tu résolus ? Précise les étapes spécifiques que tu as suivies pour préparer les données
  - 1.2. Fournis une visualisation (graphique) illustrant l'impact de la qualité des données sur les performances du modèle. Comment ces problèmes de qualité ont-ils été gérés ?

- **Question #2 : Insights et Caractéristiques (30 points)**
  - 2.1. Identifie trois "insights" pertinents liés au succès ou à l'échec des campagnes Kickstarter. Fournis une visualisation pour chaque insight.
  - 2.2. Discute des « variables de confusion (ou confondantes)[1] » qui pourraient affecter l'interprétation de ces observations. Comment les as-tu pris en compte dans ton analyse ?
  - 2.3. Comment ces "insights" pourraient être transformés en variables pour faciliter l'apprentissage d'un modèle ML ?

[1]Les variables de confusion (ou confondantes) sont des facteurs ou des variables dans une étude de recherche qui sont liés à la fois à la variable indépendante (le facteur étudié) et à la variable dépendante (le résultat), ce qui peut conduire à une interprétation fausse ou trompeuse de la relation entre la variable indépendante et la variable dépendante. Les variables de confusion peuvent créer l'illusion d'une relation de cause à effet alors que, en réalité, l'association observée est influencée par un facteur externe.

Étude de Cas

- **Question #3 : Modèle ML et Impact Commercial (30 points)**
  - 3.1. Propose une approche ML pour prédire le succès des campagnes Kickstarter en utilisant les données fournies. Explique les types de modèles, les hyperparamètres, et la validation croisée que tu utiliserais.
  - 3.2. Comment interprètes-tu les résultats produits par ta solution ML en termes de succès des campagnes ? Comment cette solution ajoute-t-elle de la valeur pour les promoteurs de projets sur Kickstarter ?
  - 3.3. Comment envisages-tu que les parties prenantes vont utiliser ta solution pour comprendre comment lancer des campagnes à haut taux de succès ? Fournis des exemples d'utilisation dans un contexte commercial.

- **Question #4 : Maintenance du Modèle (20 points)**
  - 4.1. Imaginons que ta solution est déployée et roule maintenant en production. Tu remarques que la performance de ton modèle se dégrade progressivement depuis les derniers mois. De plus, tu identifies également certaines variables dont les valeurs semblent avoir évolué durant la même période. Selon toi, quel serait une raison qui explique cette situation et comment la résoudrais-tu ?

Étude de Cas

- **Format et Soumission**
  - Renvoie tes réponses aux questions ainsi que le code que tu as produit dans un document PDF ou dans un Jupyter Notebook.
  - Assure-toi d'inclure des commentaires explicatifs dans ton code pour faciliter la compréhension et porte attention à ce que nous soyons en mesure de reproduire tes résultats. Le document doit être structuré de manière claire et inclure des graphiques pertinents.
  - Nous évaluerons ta compréhension des concepts de science des données, ta capacité à résoudre des problèmes concrets et ta créativité dans la résolution de la problématique.
  - Nous t'encourageons à aller valider directement sur le site web de **Kickstarter** ta compréhension de la signification des variables si besoin, en plus de te référer à l'annexe #1. Bonne chance !

Annexe #1 - Kickstarter

Kickstarter est une plateforme de financement participatif en ligne qui permet aux créateurs de projets de collecter des fonds auprès du grand public pour financer leurs idées, projets artistiques, innovations, produits ou initiatives. Elle a été fondée en 2009 à Brooklyn, New York, et est devenue l'une des plateformes de crowdfunding les plus populaires et les plus connues au monde.

Le fonctionnement de Kickstarter est le suivant :

- Les créateurs de projets proposent une description détaillée de leur idée ou projet sur la plateforme Kickstarter, y compris leurs objectifs financiers et le calendrier du projet.
- Les créateurs fixent un objectif de financement (un montant de fonds à atteindre) et une durée de campagne pendant laquelle les contributions peuvent être faites.
- Les personnes intéressées par le projet (les "backers") peuvent soutenir financièrement le projet en faisant des contributions, appelées "pledges". Ces contributions peuvent être de différentes natures, allant des montants symboliques aux investissements plus importants.
- Si le projet atteint ou dépasse son objectif de financement dans le délai imparti, les fonds collectés sont remis aux créateurs pour qu'ils puissent réaliser leur projet. Sinon, les contributions sont généralement remboursées aux "backers") .

Kickstarter est principalement utilisé pour financer des projets créatifs et artistiques tels que des films, de la musique, des jeux, des livres, des œuvres d'art, mais il est également utilisé pour des projets technologiques, des initiatives humanitaires, des inventions, et plus encore. La plateforme a permis à de nombreuses idées innovantes de devenir réalité en permettant aux créateurs de trouver un financement direct auprès du public, contournant ainsi les canaux de financement traditionnels.

Annexe #2 - Documentation des Données

Le jeu de données (« ks_dataset.csv ») contient les informations suivantes

Variable	Description	Type
ID	L'identifiant unique du projet correspondant. Par exemple, "1000014025".	string
name	Le nom du projet correspondant. Par exemple, "Monarch Espresso Bar".	string
main_category	La catégorie principale dans laquelle le projet s'inscrit. Par exemple, "Poésie", "Alimentation", "Musique« , etc.	string
category	Une description plus précise de la catégorie principale. Sous-groupe de la catégorie principale (voir 2.). Par exemple, "Boissons" serait un sous-groupe de la catégorie "Alimentation" de l'attribut catégorie principale.	string
currency	La devise du projet (par exemple, USD ou GBP).	float
deadline	La date limite du projet.	time
goal	Montant en monnaie locale demandé initialement par le projet	float
launched	La date de lancement du projet.	time
pledged	Montant en monnaie locale que le projet a réalisé à la date limite.	float
state	Le projet a-t-il été couronné de succès à la fin de la journée ? L'état est une variable catégorielle divisée en niveaux : succès, échec, en cours, annulé, indéfini et suspendu.	string
backers	Le nombre de supporters qui ont investi dans le projet.	int
country	Pays d'origine du projet.	string
usd pledge	Montant en USD que le projet a réalisé à la date limite.	float

Critères d’évaluations

Aptitudes Techniques

- Traitement des données : Évalue la capacité de l'équipe à nettoyer, prétraiter et transformer efficacement les données brutes.
- Feature Engineering: Évalue la créativité et l'efficacité des techniques de Feature Engineering. utilisées pour améliorer les performances du modèle.
- Sélection du modèle : Considère la pertinence des modèles d'apprentissage automatique ou statistiques choisis pour le problème et les données.
- Performance du modèle : Évalue l'exactitude prédictive du modèle, la précision, le rappel, le score F1 ou d'autres mesures pertinentes.
- Visualisation des données : Évalue la clarté et la pertinence des visualisations des données pour transmettre des informations.
- Qualité du code : Organisation, lisibilité et documentation du code.

Résolution de Problèmes

- Compréhension du Problème : Évalue à quel point l'équipe comprend le problème, ses subtilités et sa pertinence pour les scénarios du monde réel.
- Créativité et Innovation : Évalue si l'équipe a proposé des solutions ou des approches innovantes au problème.
- Robustesse : Considère à quel point la solution s'adapte aux circonstances changeantes ou aux défis supplémentaires présentés pendant la compétition.
- Débrouillardise : Évalue à quel point les candidats s'adaptent aux changements inattendus ou aux défis présentés lors du concours

Canam Group | *Artificial Intellignece*

Study Case Cameroun

November 2023

CANAM

Challenge Objectives

The main objective is to solve a data science case study. Participants have a one-week window to complete the case study and submit their deliverables. These deliverables consist of answers to the questions and the code produced in a PDF document or a Jupyter Notebook.

This is an individual task aimed at assessing the following skills:

- Fundamentals of programming
- Statistics and mathematics
- Data management
- Databases and SQL
- Data visualization
- Foundations of machine learning
- Resourcefulness and autonomy
- Problem-solving skills
- Communication / simplification skills

Study Case

- **Instructions :** Welcome to the CANAM competition, specifically to the Artificial Intelligence challenge! Your next task will be to demonstrate some of your skills through a brief data science case. We have provided you with a dataset from the Kickstarter platform (see "ks_dataset.csv" file). Using this data, we invite you to address the problem below. Once solved, please send us your answers to the questions along with the code you have produced. Good luck!

Link to the data (password: moovai) **CLICK HERE**

- **Context :** Some project promoters are trying to understand how to increase the success rate of their future campaigns. They have historical data from campaigns announced on the Kickstarter platform at their disposal.

- **Objective :** Develop a machine learning (supervised and/or unsupervised) approach in Python to assist project promoters in launching campaigns with a high potential for success.

Study Case

- **Question #1 : Data Preparation (20 points)**
  - 1.1. If you encounter data quality issues while manipulating the Kickstarter data we provided (hint: we confirm that the data was corrupted 😉), how did you resolve them? Specify the specific steps you followed to prepare the data.
  - 1.2. Provide a visualization (graph) illustrating the impact of data quality on model performance. How were these quality issues managed?

- **Question #2 : Insights and Features (30 points)**
  - 2.1. Identify three relevant "insights" related to the success or failure of Kickstarter campaigns. Provide a visualization for each insight.
  - 2.2. Discuss the confounding variables that could affect the interpretation of these observations. How did you account for them in your analysis?
  - 2.3. How could these "insights" be transformed into variables to facilitate the learning of an ML model?

[1]Confounding variables are factors or variables in a research study that are related to both the independent variable (the factor being studied) and the dependent variable (the outcome), which can lead to a false or misleading interpretation of the relationship between the independent variable and the dependent variable. Confounding variables can create the illusion of a cause-and-effect relationship when, in reality, the observed association is influenced by an external factor.

Study Case

- **Question #3 : ML Model and Business Impact (30 points)**
  - 3.1. Propose an ML approach to predict the success of Kickstarter campaigns using the provided data. Explain the types of models, hyperparameters, and cross-validation you would use.
  - 3.2. How do you interpret the results produced by your ML solution in terms of campaign success? How does this solution add value for project promoters on Kickstarter?
  - 3.3. How do you envision stakeholders using your solution to understand how to launch high-success rate campaigns? Provide examples of its use in a business context.

- **Question #4 : Model Maintenance (20 points)**
  - 4.1. Let's imagine that your solution is deployed and running in production. You notice that the model's performance has been gradually declining in recent months. Additionally, you identify certain variables whose values seem to have changed during the same period. In your opinion, what could be a reason for this situation, and how would you resolve it?

Study Case

- **Format and Submission**
  - Submit your answers to the questions along with the code you have produced in a PDF document or a Jupyter Notebook.
  - Make sure to include explanatory comments in your code to facilitate understanding and ensure that we can replicate your results. The document should be structured clearly and include relevant graphics.
  - We will assess your understanding of data science concepts, your ability to solve practical problems, and your creativity in addressing the challenge.
  - We encourage you to validate your understanding of the variable meanings directly on the **Kickstarter** website, if necessary, and refer to appendix #1. Good luck!

Appendix #1 - Kickstarter

Kickstarter is an online crowdfunding platform that enables project creators to raise funds from the general public to finance their creative ideas, artistic projects, innovations, products, or initiatives. It was founded in 2009 in Brooklyn, New York, and has become one of the most popular and well-known crowdfunding platforms in the world.

The operation of Kickstarter is as follows:

- Project creators provide a detailed description of their idea or project on the Kickstarter platform, including their financial goals and the project's timeline.
- Creators set a funding goal (an amount of funds to be reached) and a campaign duration during which contributions can be made.
- Individuals interested in the project (the "backers") can financially support the project by making contributions, called "pledges." These contributions can vary in nature, from symbolic amounts to more substantial investments.
- If the project reaches or exceeds its funding goal within the specified timeframe, the collected funds are given to the creators to carry out their project. Otherwise, contributions are typically refunded to the backers.

Kickstarter is primarily used to fund creative and artistic projects such as films, music, games, books, artwork, but it is also used for technological projects, humanitarian initiatives, inventions, and more. The platform has enabled many innovative ideas to come to life by allowing creators to secure direct funding from the public, bypassing traditional funding channels.

Appendix #2 – Data Dictionary

The dataset ("ks_dataset.csv") contains the following information.

Variable	Description	Type
ID	The unique project identifier corresponding to, for example, "1000014025."	string
name	The project name, for example, "Monarch Espresso Bar."	string
main_category	The main category in which the project falls, such as "Poetry," "Food," "Music," etc.	string
category	A more specific description of the main category. This is a sub-category of the main category (see 3). For example, "Beverages" would be a sub-category of the "Food" category in the main category attribute.	string
currency	The project's currency (e.g., USD or GBP).	float
deadline	The project's deadline.	time
goal	The initial amount requested by the project in the local currency.	float
launched	The project's launch date.	time
pledged	The amount in the local currency that the project has raised by the deadline.	float
state	Was the project successful at the end of the day? The status is a categorical variable divided into levels: success, failure, ongoing, canceled, undefined, and suspended.	string
backers	The number of backers who have invested in the project.	int
country	The project's country of origin.	string
usd pledge	The amount in USD that the project has raised by the deadline.	float

Evaluation Criteria's

Technical Skills

- Data Processing: Evaluates the team's ability to effectively clean, preprocess, and transform raw data.
- Feature Engineering: Assesses the creativity and effectiveness of feature engineering techniques used to enhance model performance.
- Model Selection: Considers the relevance of machine learning or statistical models chosen for the problem and data.
- Model Performance: Evaluates the predictive accuracy of the model, precision, recall, F1 score, or other relevant metrics.
- Data Visualization: Assesses the clarity and relevance of data visualizations for conveying information.
- Code Quality: Organization, readability, and documentation of the code.

Problem Solving

- Problem Understanding: Evaluates how well the team understands the problem, its intricacies, and its relevance to real-world scenarios.
- Creativity and Innovation: Assesses whether the team has proposed innovative solutions or approaches to the problem.
- Robustness: Considers how well the solution adapts to changing circumstances or additional challenges presented during the competition.
- Resourcefulness: Evaluates how candidates adapt to unexpected changes or challenges presented during the competition.