

Ahmed Tidiane BALDE

Ingénieur en Machine Learning

 atidiane |  ahmed-balde |  ahmedt.balde@gmail.com |  +33 6.69.78.36.06

PROFIL

Ingénieur en Machine Learning avec une expérience hybride en recherche et en consulting. En parallèle, mentor passionné par l'enseignement sur des parcours de Data Science et d'IA. Intéressé majoritairement par les domaines de la vision par ordinateur et du traitement naturel du langage en particulier par les grands modèles de langage. Apte à conduire un projet de sa conception au déploiement avec une expérience avérée en optimisation des modèles.

COMPÉTENCES

| | |
|--------------------|--|
| Analyse de données | Détection et traitement des outliers, imputation des valeurs manquantes, ingénierie des variables, visualisation des données |
| Machine Learning | Modèles linéaires, arbres de décision, modèles ensemblistes, séries temporelles, réseaux de neurones |
| Computer Vision | SIFT, CNN, ResNet, Yolo, SORT, DeepSORT, AlphaPose, OpenPose, Segmentation |
| TAL | Sac de mots, Word2Vec (Skip-gram, CBOW), TF-IDF, RNN, GRU, LSTMs, Transformers |

OUTILS

| | |
|-------------------|---|
| Langages | Python, C/C++, Java, Javascript |
| Machine Learning | pytorch, tensorflow, OpenCV, scikit-learn, pandas, numpy, scipy, prophet, transformers, peft, langchain, SHAP |
| Optimisations | deepspeed, TensorRT, onnx |
| Déploiement | docker, fastapi, mlflow |
| Tableau de bord | streamlit, gradio, dash |
| Big Data | Spark, Hadoop |
| Data Warehousing | SQL, MongoDB |
| Cloud | AWS, Azure |
| OS et versionnage | Unix/Linux, Git |

FORMATIONS

| | |
|-------------|--|
| 2017 - 2020 | Master DAC (Données, Apprentissage et Connaissance) à Sorbonne Université Sciences (Mention Bien) |
| 2015 - 2017 | Licence Mono-Informatique à l' Université de Pierre et Marie Curie (UPMC) (Mention Assez Bien) |
| 2014-2015 | Licence 1 Mathématiques-Informatique à l'Université Claude Bernard de Lyon1 (UCBL) |
| 2014 | Baccalauréat Scientifique au Lycée Sainte Marie en Guinée |

EXPÉRIENCES PROFESSIONNELLES

Machine Learning Engineer

Polarys

Feb 2024 - Présent

Colombes

Conception de chatbots

Filiale d'une grande société de transports
Pour remplacer les nombreux FAQs qu'ils avaient à disposition, le client s'est lancé dans la mise en place de chatbots basés sur le RAG (Retrieved Augmented Generation). Ce projet à haute envergure est actuellement déployé en production pour quelques centaines d'utilisateurs pour valider la stabilité du chat. Il sera par la suite adapté à de nombreux cas d'usages pour les différents départements au sein de l'entreprise. L'objectif final étant de le passer à l'échelle de centaines de milliers d'utilisateurs potentiels interne à la compagnie.

- Implémentation de la mémoire de conversation du chatbot avec la base de données Redis;
- Contribution à l'implémentation de techniques pour augmenter les performances du chatbot, entre autres : Prompt Engineering, Sentence Window Retrieval, MMR, reranking, Query Retriever;
- Supervision de stagiaires et d'alternants pour développer des stratégies de chunking avec Azure Document Intelligence et Amazon Textract;
- Amélioration de la CI/CD de l'application avec Jenkins et passage à Gitlab CI;
- Amélioration de l'environnement de travail en local (Docker) pour synchroniser les changements et la communication entre les différents containers;

Machine Learning Engineer

Pivot & Co

Dec 2022 - Fev 2024

Paris

R&D Proof-of-concepts

Internal - [voir HF](#)

Face à la complexité croissante des régulations ESG et à la nécessité pour les entreprises d'y répondre de manière efficace, nous avons identifié le besoin d'un outil interactif pour simplifier l'accès à ces informations. C'est dans cet esprit que le concept d'un chatbot ESG a été développé. Ce chatbot vise à offrir une interface conviviale et intuitive pour les utilisateurs, leur permettant de comprendre et d'interroger facilement les régulations ESG. Il s'agit d'une réponse directe à l'augmentation des exigences en matière de conformité et de transparence, offrant ainsi un moyen rapide et accessible de naviguer dans les complexités des normes ESG.

- Génération de données basées sur les régulations ESG en s'inspirant de ce qui a été fait sur *standford alpaca* sous format (Instructions Input Output);
- FineTuning de Falcon-7B sur les données générées en utilisant du Parameter Efficient Fine Tuning (PEFT) et Low Rank Adaptation (LoRA);
- Déploiement du modèle sur HuggingFace *Inference Endpoints* et développement d'un dashboard avec *streamlit* pour l'interaction;
- Même expérience réalisée avec les API d'OpenAI couplées à *langchain* pour créer un prototype de QA sur des documents sur lesquels nous avons créé des vectors stores avec ChromaDB.

Mise en place d'une pipeline de données et développement de logiciels pour la conformité aux régulations ESG

Eurazeo

Avec l'entrée en vigueur du CSRD (Corporate Sustainability Reporting Directive) en 2021, l'importance des rapports ESG (Environnemental, Social, Gouvernance) pour les entreprises a considérablement augmenté. Ces rapports, devenus obligatoires pour un nombre croissant d'entreprises, jouent un rôle crucial non seulement en matière de conformité réglementaire, mais aussi dans l'amélioration de l'image publique des entreprises. Dans ce contexte, notre client, un grand fonds d'investissement, a sollicité notre expertise pour développer un système intégré de collecte, de calcul et de reporting des métriques ESG. Ce projet visait non seulement à répondre aux nouvelles exigences réglementaires, mais aussi à positionner notre client de manière proactive face aux enjeux environnementaux, sociaux et de gouvernance. Cette démarche a renforcé la transparence et la responsabilité, éléments de plus en plus valorisés dans le monde des affaires actuel.

- Mise en place et alimentation d'un modèle de données sur Snowflake;
- Développement d'un package pour extraire les données du data warehouse, calculer les métriques ESG des différentes entreprises et fonds d'investissements que composent notre client et pousser les résultats sur Snowflake;
- Création d'un chatbot-intent pour répondre à des questions sur les législations ESG :
 - * Prétraitement des données textuelles avec NLTK, tokenisation, lemmatisation, suppression des ponctuations;
 - * Tests de différentes méthodes de *vectorisation de mots*, bag of words, puis Word2Vec (Skip-gram, CBOW);
 - * Création du Dataset du DataLoader avec Pytorch et mise en place d'un modèle de classification d'intentions constitué de quelques couches linéaires séparés par des fonctions d'activations (ReLU);
 - * Entraînement du modèle et développement de l'interactivité du Chatbot.
- Versionning des différentes étapes du projet sur Git.

Développement d'un modèle de prédiction d'attrition

GS1 France

Face à une attrition croissante des utilisateurs chez notre client, un leader mondial dans la gestion des standards d'identification unique, nous avons été sollicités pour analyser et comprendre les raisons de la résiliation des abonnements parmi ses 60 000 entreprises clientes en France. Notre mission consistait à développer un modèle de machine learning capable de prédire la probabilité de démission d'une entreprise, en utilisant des données spécifiques. Ce modèle permet aux équipes marketing de cibler et de proposer des offres personnalisées aux clients susceptibles de résilier, contribuant ainsi à la rétention et à la satisfaction clientèle.

- Analyse, prétraitement et visualisation des données clients pour comprendre les types de clients qui résilient leurs abonnements;
- Traitement des valeurs aberrantes avec une approche métier et l'approche interquartile;
- Traitement des valeurs manquantes et calculs de tests statistiques et de variance inflation factor (VIF) pour éviter les problèmes de multicolinéarité;
- Prédiction de la probabilité d'attrition pour chacun des adhérents avec des modèles d'arbres de décision tels que Xtreme Gradient Boosting *XGBoost*;
- Interprétation des modèles avec l'approche TreeExplainer de Shap pour comprendre l'importance qu'a chaque variable (feature) sur les résultats;
- Ajout de loggings pour monitorer l'exécution des scripts;
- Structuration du projet sur GitLab pour faciliter le déploiement.

Prédiction de ventes en volumes sur trois (3) années pour plusieurs marques de différents pays au sein du groupe

BEL GROUP

Après la période Covid, notre client, un acteur majeur du marché fromager, souhaitait optimiser ses investissements pour augmenter ses ventes en volume. Nous avons réalisé une étude approfondie des ventes, examinant la position de chaque marque sur leur marché respectif. En utilisant des modèles de séries temporelles développés avec la librairie Prophet, nous avons prédict les ventes sur trois ans, en intégrant divers scénarios pour évaluer l'impact de chaque paramètre comme l'Advertising & Promotion (A&P), le prix par volume, le taux d'innovation, et les promotions. Les résultats et les insights modélisés ont été présentés lors de workshops dans sept pays différents, facilitant des discussions stratégiques parmi les décisionnaires et contribuant à une meilleure compréhension des leviers d'optimisation des ventes.

- Analyse des données de ventes en volume pour comprendre la saisonnalité;
- Prédire les ventes en volume sur trois (3) années pour chacune des marques en utilisant des modèles logistiques et linéaires de la librairie Prophet, avec et sans regressors;
- Mise en place de scénarios en variant les inputs (regressors) pour ainsi prédire l'impact qu'a chaque feature dans la projection des ventes en volume;
- Préparation de workshops au sein de sept (7) pays pour présenter les résultats et expliquer nos modèles et leurs limites.

Chargé de TDs en Probabilités et Statistiques

Université Paris 1 Panthéon-Sorbonne

Fev 2024 - Présent

Paris

Ecole de Management de la Sorbonne

[Licence 2](#)

Le programme se concentrera principalement sur les aspects de dénombrement, d'introduction aux probabilités, et sur l'étude des variables aléatoires, tant discrètes que continues. J'aborderai également les concepts de densité de probabilité, les convergences en loi, ainsi que les intervalles de fluctuation et de confiance. Mon rôle consistera à traiter ses concepts avec les étudiants tout en assurant une compréhension profonde et pratique.

Mentor Data Scientist

OpenClassrooms

Fev 2023 - Présent

Paris

Accompagnement d'étudiants dans la réalisation des projets

[voir Parcours](#)

En tant que mentor en Data Science chez OpenClassrooms, une plateforme de formations en ligne, je guide les étudiants dans leurs projets pratiques. Ma tâche inclut l'assistance technique, l'explication de concepts théoriques, et la démystification des modèles et algorithmes. Je réalise un suivi hebdomadaire pour fixer des objectifs atteignables, rédige des rapports de progression, et prépare les étudiants à présenter efficacement leurs projets lors des soutenances.

- Présenter les projets et expliquer les attentes et compétences à valider;
- Guider et répondre aux questions techniques et théoriques des étudiants sur les notions de Data Science;
- Définir les objectifs à atteindre pour chaque étudiant pour la session suivante et s'assurer à respecter le délai défini en amont pour la validation de chacun des projets;
- Rédiger un rapport hebdomadaire sur leur progression;
- Valider les livrables des étudiants avant de les présenter à la soutenance.

Machine Learning and Software Engineer

Institut National de Recherche en Informatique et en Automatique (INRIA)

Mars 2021 - Nov 2021

Paris

Développement d'un système de protection piétons

[voir IPYA](#)

IPYA (I Predict Your Actions), est un projet innovant financé par l'INRIA, d'une durée de 9 mois, axé sur le développement de modèles IA pour prédire les intentions des piétons et augmenter l'autonomie des véhicules, réduisant ainsi les risques d'accidents. Ma contribution incluait l'ajustement des algorithmes de détection de piétons, et l'exploration de techniques avancées de tracking et de segmentation. J'ai également participé au développement de modèles basés sur des LSTMs et des couches de transformers pour une prédiction précise des actions futures. Le projet s'est conclu par un déploiement réussi sur Nvidia Jetson Nano pour des tests en temps réel.

- Création du Dataset et du DataLoader avec PyTorch sur la base de données JAAD (Joint Attention for Autonomous Driving) spécifiquement amélioré avec plus de labels sur ce projet;
- Etat de l'art sur les modèles de détection, de tracking, de segmentation et de prédiction d'action;
- Finetuning d'algorithmes de détections (yolov5) sur les données pour la détection de piétons;
- Exploration d'algorithmes de tracking SORT (Simple Online and Real Time Tracking) et DeepSort;
- Exploration d'algorithmes de segmentation tels que BiseNet et DDRN ainsi que d'autres qui sont plus à l'état de l'art comme HyperSeg, dans l'objectif de différencier un piéton se trouvant sur la chaussée et un autre se trouvant sur le trottoir;
- Recherche et Développement sur les modèles de prédiction afin de prédire le TTE (Time-To-Event), à titre d'exemple, le piéton va-t-il traverser dans 1, 3, ou 5 secondes;
- Développement de modèles à base de LSTMs pour prédire l'action future en prenant en entrée une fenêtre d'une demi-seconde d'actions passées (30 frames);
- Prédiction de l'action de la frame suivante, puis de celle à 1 seconde (30 frames), puis 3 secondes (90 frames) et ainsi de suite;
- Projection de la région d'intérêt (ROI) du piéton (2D) vers une dimension (1D) puis concaténation avec des données contextuelles et de l'action correspondante aux différents ROIs, et prédiction de l'action future;
- R&D de modèles à base de couches de transformers à la place de LSTMs;
- Déploiement du meilleur modèle sur Nvidia Jetson Nano, conversion de modèles en ONNX puis adaptation en TensorRT afin de faire des tests en temps réel en voiture en Île-de-France.

R&D Deep Learning Engineer Intern

Groupe PSA

Mars 2020 - Août 2020

Velizy Villacoublay

Classification d'intentions piétons par intelligence artificielle pour le véhicule autonome

- État de l'art, prise en main des données JAAD (Joint Attention for Autonomous Driving) et PIE (Pedestrian Intent Estimation);
- Récupérer les keypoints des différents piétons avec un algorithme de pose estimation notamment AlphaPose;
- Classification de piétons avec des modèles de réseaux de neurones récurrents tels que Long Short Term Memory (LSTMs);
- Exploration des transformers dans la continuité du projet;
- Présentation de papiers de recherches (RNNs, GRU) durant les workshops;
- Dépôt de brevet sur les travaux à l'institut national de propriété industrielle (INPI) sous le numéro FR3119480A1 : A method and apparatus for controlling a vehicle based on a prediction of a pedestrian's intention to cross a roadway on which said vehicle is traveling.

Prévision des retards des trains dans les transports publics bruxellois

- Créer un script pour collecter des données quotidiennement avec AWS EC2 instance;
- Analyser la situation actuelle et les tendances existantes telles que les trains ou les gares qui cumulent le plus de retards dans les différentes régions;
- Créer des visualisations claires et interactives (plot.ly et matplotlib bibliothèques) sur les cartes pour expliquer les résultats de mon analyse des données;
- Développer des modèles d'ensembles et des baselines tels qu'une simple régression linéaire pour prédire les retards avec les données prétraitées.

Prévision des salaires des joueurs de football de la Jupiler Pro League

[voir storyline](#)

- Faire du scraping (beautifulsoup4) pour recueillir des données sur les joueurs de la Jupiler Pro League;
- Fusionner les données collectées avec d'autres sources pour obtenir plus d'informations sur les joueurs;
- Nettoyer, prétraiter et prédire les salaires des joueurs;
- Faire une storyline sur les salaires dans le football avec des visualisations interactives puis y intégrer une simple régression linéaire pour prédire les salaires en fonction des caractéristiques du joueur sélectionné ainsi qu'un KNN (K-Nearest Neighbors) pour trouver les cinq joueurs Jupiler Pro League les plus proches de ces mêmes caractéristiques.

Assistant de recherche en Machine Learning

Nov 2018 - Fev 2019

Laboratoire Informatique de Paris 6 (LIP6/CNRS)

Paris

Prédition de l'affluence dans les stations de RER de l'Île de France

- Prise en main de la base de données MySQL des logs de validation des titres de transports;
- Recherche de données publiques à caractère socio-économiques sur la population de l'Île-de-France;
- Création, pré-traitement des jeux de données et application de la Backward Elimination pour la sélection de features;
- Test de différents algorithmes de régression avancée tels que eXtreme Gradient Boosting (XGBoost), Random Forest;
- Amélioration des résultats par une méthode d'Ensemble Learning, Stacking models.

BREVETS

Fouqueray, Thibault and Ahmed Tidiane Balde (Aug. 2022). "Procédé et dispositif de contrôle d'un véhicule à partir d'une prédition de l'intention d'un piéton de traverser une chaussée sur laquelle circule ledit véhicule". FR3119480A1. URL: <https://patents.google.com/patent/FR3119480A1/fr> (visited on 08/02/2023).

PROJETS

Convolutional Neural Networks

[voir sur GitHub](#)

Implémentation d'un réseau de neurones convolutifs dont l'architecture est proche d'AlexNet, pour la classification d'images sur la base CIFAR-10

Image Description with SIFT, bag of words and SVM Classification

[voir sur GitHub](#)

Dans le cadre d'un cours sur la reconnaissance des formes et l'interprétation d'images, nous avons été introduit au SIFT (Scale Invariant Feature Transform), qui est un descripteur d'image robuste à la rotation, la variation d'intensité ainsi qu'au zoom. Après avoir calculé les descripteurs de nos images distinctes, nous avons appliqué un K-Means avec un k défini à 1001 pour trouver les centres des clusters qui représentent au mieux les images. Le dernier cluster étant un vecteur de zéros pour les descripteurs vides. Puis, nous assignâmes à chacun des clusters, sa région la plus proche. Enfin, pour la tâche de classification, nous avons généré des caractéristiques pour chacune des images en fonction de notre dictionnaire visuel des régions.

Inpainting

[voir sur GitHub](#)

Dans un premier temps nous avons comparé la performance des algorithmes tels que la régression linéaire et la régression ridge sur les données USPS(United states Postal Service) pour la reconnaissance des chiffres et dans un second temps nous avons implémenté un ensemble de fonctions utilisant l'algorithme du Lasso pour prédire les pixels manquants d'une image. En un mot débruiter une image ou compléter une partie d'image manquante.

LANGUES

| | |
|------------------|----------|
| Poular, Français | Bilingue |
| Anglais | Courant |