Fernando Cañizares Romero.

Data Scientist | Machine Learning, AI and LLM specialist

(+34) 689002053
Birthday: February 6, 1995
Email: fercanrom@gmail.com
Address: Antonio Susillo 39, Sevilla, Sevilla, 41002 (España)

INTRODUCTION

From the beginning of my career, I knew I wanted to be a Data Scientist. I have always been fascinated by the application of mathematics to generate value, beyond its theoretical beauty. In the field of Data Science, I see that value tangibly reflected in the real world. I am passionate about finding meaning in data that at first glance may seem chaotic and meaningless. I am a dedicated, precise and consistent person, with a strong desire to improve every day and provide the best possible service. I value continuous learning and self-improvement.

MAIN PROGRAMMING LANGUAGES

- Python (5 years of experience)
- R (6 years NOT working experience. 1 year working experience)
- Java (2 year working experience)
- Some knowledge: C++, Rust, Golang, Scala.

COMPETENCES

- Reading, analyzing and implementing methods from scientific papers
- Python Data Science: Tensorflow, Keras, Pytorch, Tranformers (HuggingFace), Scikit-Learn and base packages like Pandas, Numpy, Scipy, Matplotlib, Seaborn
- NLP with OpenAI, GPT, Llama, LangChain, LLMs, Bert, RoBertA, XLNet, LayoutLMv3
- Computer vision with YOLO, DeepLab, U-Net, Res-Net
- Machine Learning, Deep Learning and Federated Learning
- Accelerated learning with CUDA
- APIs development with Flask and FastAPI
- Spark in Python (PySpark), SparkSQL, SparkML
- Cloud: AWS, Azure and GCP.
- MLflow, Kubeflow, DVC: Continuous integration of models y artifacts
- Git, Github, Bitbucket, Gitlab
- SQL/PLSQL (Oracle, Mysql), NoSQL (MongoDB, Elastic-Search)
- Docker
- Miscellaneous: R, R-studio and Shiny, MatLab, Java, TypeScript, React, HTML, CSS, Haskell

EXPERIENCE

DATA SCIENTIST. Machine Learning, AI and LLM specialist

***Simplekyc**, 21/01/2022 – Present*

Summary:

- Develop **Machine Learning** and **Deep Learning** models
- **Specialize in LLMs, RAGs Agents, and advanced Prompt Engineering techniques**
- **Expertise in computer vision, NLP, and machine learning models using PyTorch, Transformers, Scikit-Learn, FlairNLP, SpaCy, MLflow, DVC, and Docker.**
- **Cloud platforms: Azure** and **GCP**
- **Lead Annotation Team** using **Label Studio**

Description:

My primary role involves designing, training, and optimizing machine learning and Deep Learning algorithms, focusing on Natural Language Processing (NLP) computer vision and Large Language Models (LLMs) to real-world applications such as document understanding and information completion. I primarily use python and libraries like Pytorch, Transformers, FlairNLP and Schikit-learn to build or machine learning and deep learning models. We have dockerized every project and we manage our model's lifecycle using MLflow and we annotate our data using label studio. We have machine and resources in GCP and Azure to train our models with high computational level machines, also resources for our data, and data bases.

Key projects include text analysis using models like BERT, RoBERTa, DistilBERT, and XLNet; OCR with PaddleOCR; and handwritten text preprocessing. I have led end-to-end development of NER models using FlairNLP, XLNet, and LayoutLMv3. In document classification, I developed systems comparing models such as KNN, logistic regression, SVC, SGD classifier, Random Forest, and LightGBM, and created a voting system that achieved 98% accuracy. For document layout analysis, I explored various approaches with YOLOv8, Detectron2, and LayoutLMv3, achieving optimal performance with YOLOv8 in terms of mAP50, mAP95, recall, and processing time. Our RAG systems have been deployed for corporate information retrieval from Elastic Search indexes and generating AI agents for risk assessment.

I have also worked with vision models like YOLO, ResNET, Detectron2 and LayoutLMv3 (with a header modification) for document layout analysis to identify signatures, stamps, handwritten texts, and certifications among other classes. Collaborating closely with DevOps, product, and customer success teams, I ensure the technical solutions meet project requirements and successfully integrate AI-driven features into our products.

I created a MLflow environment using GCP machine, storage and postgresSQL data base. MLflow helps me to manage our model's lifecycle easily and to watch the training process closely on a friendly UI.

I am responsible for continuously evaluating the performance of our AI solutions, using metrics, debugging, and root cause analysis to ensure scalability, reliability, and positive impacts on user experience and business outcomes. Tools like Grafana help monitor and manage model performance in production.

Finally, I designed a labelling system using Label Studio, leveraging model predictions to pre-annotate data for review by our labelling team, enhancing our NER, layout analysis, and page classification datasets.

DATA SCIENTIST. Machine Learning, AI and LLM specialist

Redactame, 21/03/2023 – 09/10/2023

Summary:

- Developed Python API, using **Flask**, **SQLAlchemy**, **Docker**, **LLMs** and **RAGs**
- Implemented **Text Classification** and **Recommendation System**
- Cloud platforms: **Azure**

Description:

This client required an API capable of delivering accurate responses about lengthy documents, designed to function like a teacher aiding in the understanding of specific subjects. With limited resources for data handling and storage, I constructed the API using Flask, Docker, OpenAI, and Scikit-learn, deployed via Azure App Services. It extracts text from documents using OCR for image-based pages or a text size ratio method for text-based documents.

The system evaluates user queries to determine their relevance to the document, using RAG to deliver precise responses based on scored and sorted text blocks. This API has successfully distinguished between relevant and irrelevant inquiries and passed college-level examinations using the provided book information. We generalized the API to potentially integrate models beyond GPT-3.5 by replacing Python OpenAI code with Langchain, minimizing future coding and tool adjustments.

DATA SCIENTIST. Machine Learning, AI specialist

CGI, 01/10/2021-15/01/2022

Summary:

- Developed of Machine Learning and Deep Learning models, object detection and segmentation with **Tensorflow**, **Pytorch**, **Scikit-Learn** and **MLflow**, **DVC** using **Docker**
- **Computer vision** and **NLP**
- Cloud platforms: **AWS** and **Azure**
- Annotations Tools: **Label Studio**

Description:
During my experience in CGI I worked in two different project. The first one related with computer vision and the energy sector and the second one related with natural language processing and machine learning.

In the first project I worked building the architecture from some neural networks I used Tensorflow and Pytorch for image segmentation, our objective was to segment the cells of solar modules, to have them located for a posterior analysis. I built 3 different architectures, the classic U-Net, Deeplabv3 and ResUnet comparing the resultant metrics and the time performance I got the Unet model as the best one for our dataset. We used also a Tensorflow Object Detection model to locate the solar modules and then we fed with them the Unet model to locate the cells.

In the second project, I designed a machine learning training system for a classification problem and collaborated with the NLP team to develop a sentiment analysis model that synthesized information from various media sources. This model was integrated with other data sources to predict potential debt defaults using Scikit-learn and XGBoost models.

DATA SCIENTIST. Machine Learning, AI specialist

Vicomtech, 01/02/2021-15/09/2021

Summary:
- Researched cutting-edge technologies in Federated Learning, Deep Learning, and Machine Learning.
- **Specialized in computer vision and NLP using TensorFlow, PyTorch, Scikit-Learn, MLflow, Kubeflow, and DVC.**
- **Docker** expertise
- Managed **Ubuntu servers with high computation level**.

Description:
My initial role in a scientific setting involved applying my academic knowledge to real-world data challenges. I collaborated with a Greek partner to develop federated learning systems, enabling productive model training using client data while addressing legal and data sensitivity concerns.

We replicate the computational architecture that the clients wanted to use, one server with higher computational capacities as the orchestrator and three smaller servers as the data hosts.

Our projects focused on recognizing individuals in images, predicting ages, extracting text, and locating images, contributing to a Europol initiative to combat child abuse and illegal border crossings. My involvement in federated learning informed my second master's thesis, which provided an overview of the approach and tested it by training an image classification model across various machines with different weight merging algorithms.

DATA ENGINEER, JUNIOR PROGRAMMER

Fujitsu Sevilla, 15/07/2019-15/01/2021

Summary:
- **ETLs** with **Python** and **Docker** using **PySpark**, **SQLAlchemy**.
- Data Bases **Oracle**, **ElasticSearch**, **SQL**, **PL/SQL**.
- Web Maintenance, **Java**, **JavaScript**, **HTML**, **CSS**

Description:
In my first professional role, I learned fundamental skills in Python, Docker, and SQL, which have significantly benefited my current work. I was involved in a web maintenance project, creating data views for new features and resolving issues with existing ones. I used PySpark and SQLAlchemy to develop ETLs, updating client databases as needed. My work with PySpark informed my first master's thesis on its capabilities, including machine learning model deployment and real-time data processing. I also contributed to the development of new Java-based web features, enhancing user interfaces and troubleshooting web applications.

EXCITING PROJECTS

In this section, I describe some of the intriguing projects in which I have participated:

Document Reader (Simplekyc): Utilized Large Language Models (LLMs) to enhance a Named Entity Recognition (NER) system, making the backend workflow more flexible. This setup allowed us to modify just a single prompt, rather than reconfiguring the entire project to generate new models when adapting to new roles or extracting other relevant information for our clients.

Document Information Extraction (Redactame): Developed an API to extract text blocks from extensive documents. The LLM was employed to generate accurate and context-specific answers based on a comprehensive corpus of text blocks, enhancing the depth and relevance of the extracted information.

Companies' Analyzer (Simplekyc): Engineered an agent that interfaces with an API for retrieving company-specific information. This agent enhances the retrieved data with additional insights gleaned from web searches, providing a richer analysis of the company in question.

Large Document Classification (Simplekyc): Applied advanced NLP techniques to condense data size and constructed a complex training system integrating multiple machine learning models. This system was designed to identify the optimal model for effectively classifying large documents based on our specific requirements.

Solar Modules Segmentation (CGI): Designed the architecture for convolutional neural networks using models such as DeepLabv3 and U-Net, among others. The goal was to compare their performance on our test set and select the best model for integration into our API.

Risk Evaluation: Participated in two distinct projects addressing company risk assessment:

- **(CGI)** The first project utilized machine learning to classify companies based on financial details and the results of a complex sentiment analysis model that considered social media, news, and legal information.
- **(Simplelyc)** The second project involved building an Agent based on Generative AI and a search engine to classify the risks associated with the services offered by a company.

Federated Learning Viability and Tools Research (Vicomtech): Contributed to a research team exploring optimal strategies for using federated learning in addressing computer vision challenges, particularly the handling of sensitive data. We experimented with several federated learning tools, replicating the client's computational architecture, which included a primary server acting as the orchestrator and three smaller servers conducting model training. These servers then transmitted the encrypted model weights back to the orchestrator for model merging.

MLflow for Model Lifecycle Management (Vicomtech, CGI, Simplekyc): Developed MLflow environments tailored to different job requirements, deployed across various cloud providers including AWS, Azure, and GCP. I created Docker Compose files to deploy the MLflow server and user interface, integrating storage or a database as required, or utilizing pre-built resources provided by clients.

Annotation Machine Learning Backend (Simplekyc): Constructed a user interface for Label Studio and developed a machine learning backend that leverages predictions from our production models to label our dataset. If the model has previously predicted a label, the system retrieves the existing annotation, allowing the client to pre-annotate the data efficiently.

PROFESSIONAL SKILLS

Researching / Team work / Leadership / Organization and punctuality / Office / Prompt Engineering.

INTERESTS

Deep Learning / Machine Learning / Federated Learning/ NLP / Computer Vision / LLMs / Differential Privacy / Homomorphic Encryption / Mathematica / Data analytics and Data visualization / Statistic / Data Science Data Analytics / Big Data

LANGUAGES

- Spanish (Native)
- English (Professional)
- Français (débutant)

HOBBITS

Mathematics / anime / leather craft / carpentry / body building / Lord of the Rings / history / Discover new things / talk with my friends / cooking

ACADEMIC EDUCATION

- MASTER IN BIG DATA AND DATA SCIENCE
  Universidad de Sevilla (2020-2021)

- MASTER IN BIG DATA AND BUSINESS INTELLIGENCE
  INESEM BUSSINESS SCHOOL (2019)

- DEGREE IN MATTHEMATICS
  Universidad de Sevilla (2014-2019)