Ireri Avila

Email | LinkedIn | GitHub | Website | (+1) 650-398-8571 | San Francisco, CA

Data Scientist with a background in physics and 5+ years of experience, driving impactful process improvements in industry and academia.

Education

M.S in Data Science — ***Jul 2023 – Jun 2024***
University of San Francisco — *San Francisco, CA*
Relevant Coursework: Data Acquisition, Machine Learning, Regression, Time Series, A/B Testing, Relational Databases (SQL), NoSQL

B.S in Physics — ***Aug 2016 – May 2020***
Universidad de las Américas Puebla — *Puebla, MEXICO*
Relevant Coursework: Object-Oriented Programming, Matrix Theory, Probability and Statistics, Computer Simulation

Professional Experience

Data Scientist — ***October 2023 - Present***
Give Us The Floor — *Remote, US*

- Designed and deployed ML models leveraging NLP for binary classification, obtaining under 1% Type II errors on unbalanced datasets.
- **Reduced human intervention by 90% saving $90K per year** and impacting more than 2,000 users.
- Improved message retrieval through API requests by 80% with a Python script for effective date filtering for 400K messages.
- Engineered an Apache Airflow DAG to automate end-to-end pipeline runs, streamlining data fetching, processing, and final classification, enhancing efficiency and scalability; reducing human interaction time by 60%.
- Developed comprehensive onboarding materials and structured curriculum for incoming data science interns; improved intern performance metrics by reducing time-to-competence from one month to two weeks.

Data Analyst, Internal Global Audit — ***Mar 2021 – May 2022***
Grupo Bimbo — *Mexico City, MEXICO*

- Creation of **15 analytics scripts** for audit reports from an Oracle Database connection
- Implementation of Python scripts for the automatization of monthly programmed tasks, **saving over 3 hours per month.**
- Worked closely with cross-functional teams to create a Dynamic Risk Assessment dashboard for fixed assets using Power BI, reducing reporting time by 70% across 5 organizations..
- Validated data consistency across 3 Azure data lakes containing over 2 million records, ensuring data integrity through SQL queries.
- Optimized and refined scripts that previously crashed for large inputs of data while reducing **computing time by 40%.**

Data Scientist, Research — ***May 2019 – May 2020***
National Institute of Astrophysics, Optics, and Electronics (INAOE) — *Puebla, MEXICO*

- Data mining of around 250K records of data from 28 different sources using Beautiful Soup.
- Implemented a Python script for data normalization, analysis, and time series modeling, reducing data preprocessing time by 30%.
- Created data visualizations using Matplotlib and Seaborn, transforming complex datasets into clear, interpretable insights, leading to a 25% reduction in analysis time.

Projects

AI Generated Podcasts - Google AI Hackathon - ***May 2024***

- Podcast generation of research papers by interactions between two LLMs..
- Built Python scripts to integrate with Google Gemini's API, retrieving and processing Q&A data from documents, and generating conversational transcripts using advanced prompt engineering techniques.
- Designed code to connect with Google's Text-to-Speech API. Generated podcast audio for both LLM speakers using Google's Text-to-Speech API.

Group Listening Dynamics with Spotify API

- Executed project's pipeline in Airflow through Google Cloud Composer as well as deployment in Google Cloud Platform for data retrieval from Spotify's API for a group of users, keeping over 5K data points in a MongoDB cluster for later processing through SparkSQL.
- Co-authored an insightful Medium post showcasing the end-to-end data pipeline and visualization results, facilitating knowledge sharing within the analytics community and garnering over 100 views.

Color Analysis with AI

- Obtain tonality and seasonality based on pictures, using neural networks for multiclass classification with 85% accuracy..
- Mined and curated dataset for all 12 categories for color analysis, over 10K images..
- Engineered neural network with triplet loss for embeddings, optimizing performance for downstream tasks by 15%.

Skills

- Python, SQL, NoSQL, MongoDB, DataBricks, Transformers, HuggingFace, Google Cloud Platform, AWS, A/B Testing, Airflow, Spark, PyTorch, NLP, Oracle, Git, GitHub, Spacy, Plotly, Tableau, PowerBI, ACL Analytics, LaTeX, English, Spanish, French, Italian