

Fred Mubang, PhD

fredmubang@gmail.com

Data Scientist / ML Scientist / AI Scientist

Linkedin: Fred Mubang
Website: fmubang.github.io

Summary: I am a PhD-level Senior Data Scientist with 7 years of experience. I am currently working as a Senior Data Scientist at Experian. In addition to that, I am the Lead Developer and Data Scientist of Finvalia, an AI-driven financial newsfeed and analytics platform. I was previously a Machine Learning Researcher and Data Scientist the Department of Defense/DARPA. I have a strong background with data and statistical analysis, as well as with machine learning algorithms. My strongest technologies are Python, Pyspark, SQL, Pandas, Large Language Models, Neural Networks, and XGBoost.

WORK EXPERIENCE

Lead Developer / Data Scientist at Finvalia - [Link](#)

July 2023 - Present

- Developed and managed Finvalia, an AI-driven financial newsfeed and data analytics platform, leveraging a Large Language Model (LLM) for sentiment analysis on stock-related articles.
- Designed and implemented Python/Flask back-end systems for querying LLMs, retrieving data via APIs, and calculating hypothetical profit and loss scenarios for traders.
- Oversaw the end-to-end design of the website, focusing on user experience and functionality.
- Led the recruitment and management of web developers and UI designers, establishing project roadmaps and ensuring timely delivery of milestones.

Senior Data Scientist at Experian

Oct 2022 - Present

- Built an internal Large Language Model (LLM) chatbot using Transformer neural networks, enhancing data accessibility for thousands of Experian employees.
- Enhanced algorithms for identifying the best contact information for 200MM individuals, achieving a 30% lift in phone prediction and 20% in email prediction.
- Developed a household size prediction algorithm using Linear Regression and SMOTE, saving the organization over \$20 million in potential revenue loss.
- Leveraged Python, Pyspark, and Hadoop to perform advanced data analysis and model development, leading to actionable business insights.

Machine Learning Researcher / Data Scientist for Department of Defense/DARPA - [Link](#)

Oct 2017 - Oct 2022

- Created machine learning models (XGBoost, neural networks, LSTMs) to predict social media activity for 50 million users, aiding in the detection of misinformation spread.
- Developed and fine-tuned BERT models for topic labeling across various social media platforms, resulting in a 20% lift over historical baselines.
- Conducted comprehensive data engineering tasks, including data cleaning, feature engineering, and visualization, using Python and related libraries.

EDUCATION

Ph.D in Computer Science, University of South Florida, GPA: 3.87/4.00

Aug 2018 — Oct 2022

Master of Science, Computer Science, University of South Florida, GPA:3.87/4.00

Aug 2018 — May 2021

Post Bachelor Studies, Computer Science, University of South Florida, GPA:3.7/4.00

Aug 2017 — July 2018

Bachelor of Arts, Music Business, Berklee College of Music

Aug 2010 — May 2014

Relevant Courses: Data Mining, Machine Learning, Neural Networks, Advanced Neural Networks, Social Media Mining, Network Science, Natural Language Processing, Intro to AI, Calculus 1-3, Linear Algebra, Probability and Statistics

SKILLS AND TECHNOLOGIES

- **General ML Skills:** Building ML models from scratch, Deep Learning, Neural Networks, Large Language Models (LLMs), ChatGPT, BERT, Transformers, Recurrent Neural Networks (RNNs), Long-Short Term Memory Networks (LSTMs), Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Time Series Forecasting, XGBoost, Gradient Boosting, Probabilistic Modelling, Classification, Regression, Clustering, NLP, Word2Vec, Node2Vec, Feature engineering, Dimension reduction techniques, Logistic Regression, Linear Regression, Data Augmentation, SMOTE
- **Data Analytics:** Descriptive Statistics, Cleaning, Manipulation, Scraping, Visualization, Financial Modeling, Data Analysis
- **Business Skills:** Communicating Analysis, Leading group discussions, Mentoring, Presenting Results Clearly to Stakeholders, Leading Business and Technical projects, Developing Analytical Solutions
- **Technologies:** Python, C, C++, Linux, Scikit-learn, Pandas, SQL, Pyspark, Hadoop, Tensorflow, Keras, XGBoost, Numpy, Seaborn, Matplotlib, Networkx, Jupyter Notebooks, Microsoft, Microsoft Powerpoint, Excel, Microsoft Office