AMIT VERMA

Lead Data Engineer cum Scientist
(Orange Business Services)

CONTACT INFORMATION

Phone: +91 9811-097-827
Email: amitverma.dce@gmail.com
Location: Rohini, Delhi, India

SOCIAL

LinkedIn: amit-verma-240992
Github: git-amit-verma

CAREER SUMMARY

With 9+ years of experience, I specialize in **machine learning, deep learning, and Generative AI solutions, including prompt engineering and Retrieval-Augmented Generation (RAG) techniques**. I have successfully developed and fine-tuned transformer models for text analytics and generative tasks, resulting in significant improvements in model accuracy and responsiveness. I have worked on various classical machine learning use cases involving classification and regression, with an in-depth understanding of the mathematical concepts behind each classical ML algorithm, which helps me apply established techniques to solve real-world problems.

My expertise also extends to big data technologies such as **Hadoop, Oozie, Spark, YARN, Nifi, HBase, and Hive,** enabling me to effectively manage and process large-scale datasets. I have worked on different **connectors (scripts)** to perform various ingestion-related tasks, including cleaning and enrichment modules, which carry out the **ELT process—extracting, loading, and transforming data** into optimized formats—ensuring seamless data integration and processing. Additionally, I have solid development experience in building web applications with Python frameworks, and I am familiar with tools such as Spotfire and QlikSense.

Currently, **I am leading two separate teams:** one focused on supporting **ingestion-related requests** and assignments managed through JSM or other ticketing tools, and another **team dedicated to Generative AI,** where we are incorporating Large Language Model (LLM)-based solutions into existing applications.

KEY SKILLS

- **Machine Learning & Deep Learning:** Expertise in developing and fine-tuning machine learning and deep learning models. Experience with classical machine learning use cases, focusing on classification and regression tasks.
- **Generative AI:** Proficient in designing and developing advanced Generative AI solutions, leveraging prompt engineering techniques and implementing Retrieval-Augmented Generation (RAG) to enhance model outputs with real-time, contextually relevant data. Skilled in fine-tuning large language models (LLMs) for tasks such as text generation, summarization, and multi-turn conversations, resulting in more accurate and responsive AI-driven applications.
- **Natural Language Processing (NLP):** Skilled in using transformer-based models for text analytics and generative tasks.
- **Big Data Technologies:** Proficient in technologies such as Hadoop, Oozie, Spark, YARN, Nifi, HBase, Hive for managing and processing large-scale datasets.
- **Data Engineering**: Worked on different connectors (scripts) for ingestion-related tasks. Knowledge of the ELT process (Extract, Load, Transform) for data optimization.
- **Web Development:** Good development experience in building web applications and APIs using Python frameworks.
- Knowledge of **Dockers and Kubernetes** for deploying and managing applications in a containerized environment.
- Familiarity with **Data Visualization** Tools such as Tableau, Spotfire and QlikSense.
- **Collaboration Tools**: Experience with tools like JIRA for project management and team communication.

TOOLS/FRAMEWORK

- Programming Language: Python (3. x version), Go Lang
- Data Pre-processing: Pandas, NumPy, PySpark
- Machine/Deep Learning Frameworks: Scikit-Learn, Pytorch, Transformers
- Generative AI: LangChain, LangSmith, and LangGraph modules for streamline AI-driven workflows, integrating with cloud solutions like CloudGroq for free LLMs.
- Natural Language Processing: NLTK, Spacy, Transformers
- Big Data Technologies: Apache Hadoop, Spark, Kafka, Nifi, Oozie, Hive
- Containerization and Orchestration: Docker and Kubernetes
- Development: Django, Rest-Framework, FastAPI
- Databases: SQL, Hive
- Data Modeling & Visualization: Matplotlib, Seaborn, Plotly, Spotfire, & QlikSense
- Collaboration and Experiment Tracking: Jupyter Notebooks, Google Colab, Kaggle Kernel.
- Integrated Development Environments: VS code, PyCharm, SQL Server Management Studio, GoLand, etc.
- Miscellaneous: Git for version control, JIRA for Agile Software Development

WORK EXPERIENCE

- **Role: Lead Data Engineer cum Scientist(April'24 - Present)**
  Company: Orange Business Services, Gurugram, Haryana, India

  Description:
  Currently, **leading two separate teams:** one focused on supporting **ingestion-related requests** and assignments managed through JSM or other ticketing tools, and another **team dedicated to Generative AI,** where we are incorporating Large Language Model (LLM)-based solutions into existing applications.

  UC Support1(Use-case support India Team): Ingestion team(9 people) at EDH (Enterprise Data Hub) in India, a centralized data solution for the entire Orange organization. We ingest data from various business/sources after obtaining approval from business stakeholders and integrate it into our Big Data framework. The data ingested is then cleaned, enriched, and exported based on requests for data analysis and other needs.
  Technical Specifications:
  - Batch Processing: Using Apache Hadoop framework and Apache Spark for processing large volume of data.
  - Stream processing: Using Apache Kafka for real-time data processing.

  Generative AI - Culturation: Core Generative AI team (3 members) within EDH India, where we have developed two projects:
  - **SM-Incident Reporting Toolbox(In-prod)**: Developed a solution that fetches trouble ticket data from an API engine to generate **executive summaries and sequence-to-sequence summaries**. This implementation utilizes Azure OpenAI with Retrieval-Augmented Generation (RAG) techniques, where we have configured specific business-related system messages and user prompts to enhance the relevance and accuracy of the generated outputs.
  - **Dinotoo application:** Contributed to the development of the Dinotoo application, which integrates existing Azure OpenAI models and various free large language models (LLMs) such as Mistral into a unified platform, similar to ChatGPT. Leveraging LangChain, we have created a seamless interface that allows users to harness the capabilities of multiple LLMs for diverse applications, enhancing accessibility and functionality while delivering an interactive and user-friendly experience.

  Additionally, I delivered a podcast globally within the entire Orange Data & AI community on the topic "from history to modern AI".

- **Role:** Business Process-Advanced(Data Scientist)(**November'22 - April"24**)
  Company: Agilent Technologies, Manesar, Gurugram, Haryana, India

  ### PROJECT: AGILENT CUSTOMER EXPERIENCE
  Description: User experience portal for predicting the sentiment, multilabel and text summarization.
  Technical Specifications:
  - Application: Frontend & Backend with Django.
  - Machine Learning: 3 Transformer-based models leveraging BERT both for tokenization and prediction.
  - API: ML API built using FASTAPI

- **Role:** Machine Learning Engineer- Senior Software Engg.(**December'20 - October'22**)
  Company: Orange Business Services, Gurugram, Haryana, India

  ### PROJECT: NECTAR
  Description: Predict future incidents and cater to flapping issues.
  - Build a script for capturing router logs from multiple devices.
  - Created a logic to extract True flapping scenarios by analyzing device logs (CPU, router, bandwidth, memory).
  - Extract, Transform, and Load performed on historical incident data to find issue type per device.
  - Build an ML model using XG Boost with an accuracy of 79% and output issue type based on the variety of historical events.

  ### PROJECT: CUSTOM & TRADE- AUDIT AUTOMATION
  Description: It is an Artificial Intelligence-based audit application built to perform auditing of the customs and trade modules (Indirect Tax, Asset Transfer, and export Import Compliance). Application leverage "ABBY OCR" for extracting the texts from the uploaded documents (Proof of Documents, airway bills, etc.). The extracted text is fed into the ML model for document classification.
  Technical Specifications:
  1. Web-framework: PHP
  2. Database: SQL DB
  3. ML Model: Balanced Random Forest

  ### PROJECT: DATA QUALITY INDEX
  Description: It is a web-based portal to enrich the data of RFO (Reason for Outage) reports based on ML algorithms. The algorithms, along with exception handling, process and refine the RFO data. The portal has a GUI for users to upload and view the processing and downloading of the reports.
  Technical Specifications:
  1. Web-framework: Django
  2. Database: MongoDB (no SQL DB)
  3. ML Models: 3 deep learning models were created for finding Severity, RFO, and Problem Family.

- **Role:** Data Analyst**(April'17 - June'20)**

  Company: Indian Renewable Energy Development Agency of India - Contractual, Delhi, India

PROJECT: NET ENERGY GENERATION AND COSTING

- Created machine learning models with Python and Scikit learn to predict Net green energy generation based on the Joint Meter Readings (JMR).
- Designed a machine learning model for calculation in the percentage depreciation value of commissioned machinery.
- Used different Feature Selection methods for the model, maintaining the model with the new guidelines as per the changes in the Generation Based Incentive scheme.
- Created customized reports in Tableau for data visualization.

PROJECT: LOAN DISBURSEMENT

- Designed models based on the reports of loans disbursed.
- Analyze large datasets of data to provide strategic direction.
- Used different Feature Selection methods for the model, maintaining the model with the new customer trend.

- **Role:** Technical Engineer(December'14 - March'17)

  Company: IT Cons e-Solution PVT LTD
  - Ensured that all production databases are running efficiently 24/7
  - Implemented appropriate backup and restore strategies to protect all data assets.
  - Monitored database performance, tracked and stored procedures and queries' execution times, and implemented efficiency improvements.
  - Performs the following database administration tasks such as DBMS installations and upgrades, migrating databases, changing server configuration parameters, transferring system databases, and managing users and logins
  - Prioritizes assigned tasks and ensures completion on or before the scheduled date.

CERTIFICATIONS

- Data Visualization with QlikSense
- Data Modeling with QlikSense
- **Natural Language Processing (Specialization by DeepLearning.AI) :**
  1. Natural Language Processing with Classification, and Vector Spaces.
  2. Natural Language Processing with Probabilistic Models
  3. Natural Language Processing with Sequence Models
  4. Natural Language Processing with Attention Models
- **Programming with Google Go (Coursera)**: **Go Specialization**
  1. Getting started with Go
  2. Functions, Methods, and Interfaces in Go
  3. Concurrency in Go
- **Machine Learning by Stanford University:** 11 weeks course by andrew-ng.
- **University of Michigan (Coursera): Python Specialization**
  1. Programming for Everybody
  2. Using Databases with Python
  3. Using Python to Access Web Data
  4. Data with PythonPython Data Structures
  5. Capstone: Retrieving, Processing, and Visualizing
- **Imperial College of London (Coursera):** Mathematics for Machine Learning (Linear Algebra)
- **Koenig Solutions** (Google authorized partner): 7 Day workshop on Android Development Life cycle.
- **Wipro Certified Engineer (2016).**
- **International English Language Testing System (IELTS) - Academic** Listening: 7.5, Reading: 6.0, Writing: 6.5, Speaking: 7.0, Overall Band: 6.5

RESEARCH PAPER

IEEE: Fuzzy C-means with non-extensive entropy regularization

Published in: 2015 IEEE International Conference on Signal Processing, Informatics, Communication, and Energy Systems (SPICES)
Date of Conference: 19-21 Feb. 2015
Abstract: A new fuzzy c-means clustering with non-extensive entropy regularization is proposed in the paper. The purpose of entropy regularization is to form approximate solutions to singular problems in the maximum entropy framework. The non-extensive entropy with Gaussian gain is generally used for identifying non-uniform probability densities as in regular texture patterns. It is thus well suited for regularizing the FCM problem due to the presence of extremal points in real-world datasets which translate to uneven probability graphs. The new objective function is formulated and the update equations are derived subject to the constraint which is the same as that of fuzzy c-means clustering. The result is a highly improved clustering accuracy superior to state-of-the-art methods when tested on benchmark UCI datasets.

EDUCATION

- **DELHI TECHNOLOGICAL UNIVERSITY (DTU) :** B.Tech in Information Technology (2010-2014)
- Senior Secondary: CBSE (2009-2010)
- Matriculation: CBSE (2007-2008)