Karan Yadav

karanyadav.career@gmail.com | +91 9284966566
karan-yadav | karanyadav14

Profile

Result-driven Data Scientist skilled in deploying scalable machine learning models in cloud environment such as AWS, with strong foundation in Python, SQL, Statistical Data Analysis and MLOps. Experienced in building robust data pipelines leveraging Big Data technologies like Apache Hadoop and Spark. Passionate about utilizing data to drive strategic insights and facilitate informed decision-making.

Skills

Languages: Python, SQL.

Libraries: Pandas, Scikit-Learn, LightGBM, XGBoost, TensorFlow, PyTorch, Pycaret, Transformers, LangChain.

Cloud: AWS - Sagemaker, Lambda, EC2, S3, Timestream, Iot Core, SQS, SNS, EMR, ECR, ECS, MWAA.

CI/CD: GitHub actions, Docker, Terraform.

Work Experience

Tvarit GmbH, Pune Jul 2022 - Present

Data Scientist

- **Saved 20% costs** for automobile wheel manufacturers by scaling and maintaining **Prescriptive Analytics pipeline** backed by **AWS SageMaker**.
- **Reduced cost** of Predictive Analytics pipeline **by 98%** through optimized data ingestion leveraging **AWS EMR**.
- Optimized **Extract, Transform and Load (ETL) pipelines** using **AWS ECR, ECS and Apache Airflow**.
- Performed **Exploratory Data Analysis (EDA)** using **Pandas** to uncover patterns in LPDC time series data, directly contributing key features to the product.

Education

IIT Delhi Aug 2020 - May 2022
MSc, Cognitive Science ***CGPA: 8.80/10***
Relevant Coursework: Machine Learning, Probability and Statistics, Advanced Data Analysis using R.

Government Engineering College, Aurangabad Aug 2014 - May 2018
BE, Electrical Engineering ***CGPA: 7.29/10***
Relevant Coursework: Signal Processing, MATLAB, Digital Electronics.

Projects

- **LLaMA-Based Conversational AI Chatbot (2024):**
  - Developed an **LLaMA2-based chatbot** using **LangChain** and **Streamlit** to enable querying of local and Google Drive files.
- **Understanding Disfluencies in Hindi Dialogue: A Psycholinguistic Analysis (2022):**
  - Created Hindi spontaneous dialogue corpus from **24** unscripted telephonic conversations, consisting of **23947** sentences and **152720** tokens.
  - Conducted **hypothesis testing** using **Generalized Linear Mixed Models (GLMM)** in **R** to analyze the impact of unigram/bigram frequencies and dependency length on disfluencies in Hindi speech production.
- **Classification of Motor Imagery based on Electroencephalography (EEG) data (2021):**
  - Classified motor imagery of motor cortex based on EEG data using machine learning models **(Support Vector Machines, Random Forest, Decision Tree)** with average accuracy of **98.47%** for manual method and **99.12%** for autoencoder method.

Certifications

- Machine Learning A-Z on Udemy.
- Ultimate AWS Certified Cloud Practitioner CLF-C02 on Udemy.
- Data Analysis with Pandas and Python on Udemy.