Ankush Arora

Date of birth: 23/11/1995 **Phone number:** (+49) 1777056492

Email address: ankusharora.2311@gmail.com

WhatsApp Messenger: +49-1777056492

LinkedIn: www.linkedin.com/in/ankusharora23/

Home: Neustrasse 5, 56072 Koblenz (Germany)

ABOUT ME

A Machine Learning and Data Analytics enthusiast with several years of industry experience in development. Expertise in Fine Tuning LLM's, Developing CI/CD pipelines, Data Structures, ETL/ELT and Python. Proven track record of delivering process optimization and predictive analytics solutions while collaborating with cross-functional teams in Agile environments. Want to utilize my skills and abilities in favor of organizational objectives and to make myself a mark of recognition in the organization.

WORK EXPERIENCE

Data Engineer

Capgemini [16/07/2017 – 15/08/2020]

City: Mumbai | **Country:** India

Bayer AG (Business Intelligence Solution)

- Led a team of 5 data engineers in designing and implementing a robust data integration and business intelligence solution, optimizing pharmaceutical sales reporting and operational processes, while ensuring adherence to project timelines and quality standards.
- Developed high-performance ETL pipelines using IBM DataStage, Talend, MySQL, and Oracle, increasing data processing speed by 30%.
- Managed CI/CD pipelines, ensuring the smooth deployment of updates, and provided ongoing technical support to enhance system scalability and reliability.
- Utilized JIRA for issue tracking, sprint planning, and collaboration, ensuring effective project management and timely delivery of solutions.

Vertiv (Enterprise Data Warehouse Solution)

- Designed and implemented a comprehensive data warehouse solution that integrated and consolidated data from multiple sources to support enterprise-level business intelligence.
- Applied advanced SQL queries to optimize ETL workflows, resulting in a 20% reduction in report generation time and improved data accessibility for stakeholders.
- Developed and maintained interactive dashboards in Tableau, enabling more efficient analysis of business performance and key metrics.

Daimler Overseas (Data Integration and Reporting Solution for German Market)

- Managed data integration and reporting systems, specifically focused on the German market, ensuring data accuracy and consistency across multiple sources.
- Led the design and optimization of ETL processes to consolidate sales, operational, and financial data, enabling better decision-making for Daimler's overseas operations.
- Developed interactive dashboards, providing real-time insights into market performance, sales trends, and operational efficiency for the German market.

PROJECTS

[10/07/2024 – 10/02/2025]

Detecting Online Abuse: Fine-tuning LLMs for Abusive Language Detection

- Developed an advanced system for detecting online abuse by fine-tuning a Large Language Model, achieving an accuracy of 85% across 6 abuse categories including age, ethnicity, gender and religion.
- Integrated contextual embeddings with sentiment analysis features to enhance the detection of subtle forms of abuse, demonstrating significant improvements over traditional BERT and RoBERTa models.

- Utilized a dataset of 47,000 annotated tweets and 50,000 IMDB reviews for sentiment analysis to enhance the detection framework, employing techniques such as hyperparameter optimization and transfer learning.
- Implemented a custom neural network architecture combining DistilBERT's contextual understanding with sentiment scores, enabling the system to capture both semantic meaning and emotional undertones in text.

Link: github.com/ankusharora23/online-abuse-detection

[08/09/2024 – 08/11/2024]

NYC Yellow Taxi Data Pipeline

- Created a deployment ready end-to-end data pipeline in Python to process NYC Yellow Taxi trip data, implementing automated ETL workflows for data cleaning, monthly average calculation, and rolling window analysis.
- Designed and implemented an SQLite database schema to efficiently store and manage taxi trip metrics, utilizing pandas for data transformation and SQL for persistent storage of both historical and current data.
- Developed a scalable architecture with detailed documentation for distributed processing using Apache Spark/ Dask, enabling the pipeline to handle increasing data volumes through horizontal scaling and cloud storage integration.
- Built a CLI tool with configurable parameters for year, month, and rolling window size with RestAPI, incorporating comprehensive testing suite and error handling.

Link: github.com/ankusharora23/yellow-taxi-data-analysis

[01/08/2023 – 01/03/2024]

A Journey into the Future: An Applied Exploration of Time Series Forecasting (Collaborative Research and Blog Publication)

- Published a collaborative blog, *"A Journey into the Future: An Applied Exploration of Time Series Forecasting,"* on the Databricks Community Technical Blog.
- Conducted research with the University of Koblenz and Databricks, examining time series forecasting applications in retail, finance, and IoT.
- Benchmarked forecasting algorithms, including ARIMA, SARIMA, LSTM, XGBoost, and Prophet.
- Led a team of 4 researchers, providing data-driven insights and practical guidance for forecasting practitioners.

Link: community.databricks.com/t5/technical-blog/a-journey-into-the-future-an-applied-exploration-of-time-series/ba-p/55494

[12/01/2023 – 15/04/2023]

Pseudo Contraction in Knowledge Dynamics in AI (Research and Seminar)

- Researched belief changes and pseudo contraction theories in artificial intelligence, focusing on evolving knowledge bases and handling contradictory information.
- Analyzed the impact of the AGM postulates on AI systems and proposed methodologies to improve adaptability in belief-changing processes.

EDUCATION AND TRAINING

MS Web and Data Science

University of Koblenz [01/11/2021 – 10/02/2025]

Address: Universitätsstraße 1, 56070 Koblenz (Germany) | **Website:** https://www.uni-koblenz-landau.de/de/

Bachelors of Technology Computer Science

Dr. A.P.J. Abdul Kalam Technical University [30/06/2013 – 31/05/2017]

Address: G.L. Bajaj Institute of Technology and Management, Plot No, 2, APJ Abdul Kalam Rd, Knowledge Park III, Greater Noida, Uttar Pradesh, 2013006 Greater Noida (India) | **Website:** https://aktu.ac.in

DIGITAL SKILLS

Programming and Development

SSIS, SSAS, SSRS / JIRA / Talend Open Studio for Data Integration (ETL) / Prophet / Large Language Models / SARIMA / I BM Datastage / ETL and ELT / Git / MySQL / Python (Tensorflow, Keras and Pytorch) / CI/CD / Snowflake Data warehouse / Agile (Scrum) / Cloudera Impala

Data Science and Machine Learning

RESTful Webservices / CUDA (GPU) / Azure Databricks / Python / Data Science & Data Analytics / Hyperopt / Time Series Prediction / Optuna

Visualization & Analytics

Data Visualization (Matplotlib ,Seaborn, Plotly) / Data Visualization (Tableau, Power BI)

LANGUAGE SKILLS

Mother tongue(s): Hindi

Other language(s):

English

LISTENING C2 **READING** C2 **WRITING** C2

SPOKEN PRODUCTION C2 **SPOKEN INTERACTION** C2

German

LISTENING A2 **READING** A2 **WRITING** A2

SPOKEN PRODUCTION A2 **SPOKEN INTERACTION** A2

Levels: A1 and A2: Basic user; B1 and B2: Independent user; C1 and C2: Proficient user

CERTIFICATIONS

Snowflake Hands-On Essentials: Data Warehouse

Link: https://achieve.snowflake.com/a6be267d-1f1b-4081-a9d0-e9afe9ee2c35

Python for Data Science & Machine Learning Bootcamp (Udemy)

Link: http://ude.my/UC-dec55292-8b38-432-a554-9ab8b80939ec

NLP- Natural Language Processing using Python (Udemy)

Link: http://ude.my/UC-e0c082a2-afdd-4204-bac1-9ebc9ea99eb9