Sarthak Vajpayee

Dallas, TX-75252 | +1(469)-347-9198 | sarthak.vajpayee05@gmail.com | LinkedIn | Github

SUMMARY

Accomplished Data Scientist with 5+ years of experience in advanced analytics, machine learning, and cloud computing. Proven in optimizing model performance, reducing error rates, and enhancing efficiency. Skilled in leveraging cutting-edge technologies to drive innovation and deliver impactful insights. Seeking to apply expertise in dynamic tech environments to achieve strategic goals.

EDUCATION

The University of Texas at Dallas, Texas **Aug 2022 - May 2024**
Master of Science - Business Analytics (Specialization in Data Science)

Dr. A.P.J. Abdul Kalam Technical University, India **Aug 2014 - Jun 2018**
Bachelor of Technology - Electronics & Communication Engineering

PROFESSIONAL EXPERIENCE

University of Texas at Dallas – Richardson, Texas, USA **Jan 2024 - May 2024**
Data Science Research Assistant

- Analyzed and synthesized 80+ studies on prompt engineering to optimize LLMs, evaluated the effects of different prompt techniques on effectiveness and responsiveness, and harnessed the ChromaDB database for detailed data analysis.
- Employed PEFT techniques including QLoRA and IA3 in Python with Huggingface to optimize Transformer models like BERT, DistilBERT, and RoBERTa, boosting the F1 micro-score for a classification task by 40%.
- Developed ETL flows for semantic search models, boosting data throughput by 25% and improving operational efficiency.

AppSteer – Dallas, Texas, USA **May 2023 - Dec 2023**
Data Scientist (Full-Time CPT)

- Streamlined development of prompt engineering for LLMs, reducing error rates through collaboration with cross-functional teams utilizing Python, Langchain, Huggingface, PyTorch, and FAISS vector database.
- Developed automation tools using Docker, Flask, Python, and OpenAI APIs on Azure, integrating PySpark for efficient dataflow, and PostgreSQL, reducing app development and deployment time from 4 days to under 2 minutes.
- Orchestrated over 50 cloud ETL deployments, enhancing CI/CD workflows with Git version control, Jenkins, Ansible, Kubernetes, Prometheus, Grafana to boost the deployment rate by 20%.
- Led REST API development with Pydantic, FastAPI, and Python, successfully delivering features within six weeks.

Ernst & Young – Bengaluru, India **Dec 2021 - Jul 2022**
Staff Data Scientist

- Enhanced demand forecasting for a top FMCG company using XGBoost and Regression models, achieving 8% MAPE.
- Utilized Python, Pandas, and Numpy for data manipulation, and Airflow for ETL, reducing data processing times by 40%.
- Engineered a second-generation predictive system, integrating ARIMA and XGBoost in PySpark for data modeling, which achieved 10% MAPE within 15 days, setting new benchmarks for international market analytics.
- Identified critical demand trends through time-series analysis and A/B testing, resulting in a 15% increase in forecast accuracy, and leveraged Tableau dashboards for enhanced data visualization.
- Employed Azure Databricks for data processing and Azure ML for feature engineering and data mining, enhancing sales forecast models by 20% and yielding $1.1M in annual cost savings.

Scaler – Hyderabad, India **Sep 2018 - Nov 2021**
Data Science Engineer

- Architected a robust LMS and online coding platform using FastAPI, Kubernetes, AWS, and Docker for containerization, integrating an NLP-based ticketing system with Python, boosting resolution efficiency by 25%.
- Developed an end-to-end plagiarism detection tool with Doc2Vec, FastText, and BERT using Python and TensorFlow.
- Engineered and optimized CI/CD workflows using configuration management tools such as Ansible, integrated with Jenkins, leading to a 20% increase in deployment frequency and a 30% reduction in lead time.

PROJECT EXPERIENCE

CI/CD for Credit Risk Prediction

- Developed and deployed an ETL pipeline and Streamlit app for Credit Risk Analysis using Python, Decision Trees, PostgreSQL, Jenkins, Docker, enabling 20% faster credit decisions and improved efficiency through CI/CD.

SKILLS

Programming: Python, R, SQL, NoSQL, MongoDB, JavaScript, C, MATLAB, Bash, SAS, Linux/Unix
Libraries: Scikit-Learn, Spacy, NumPy, Pandas, Keras, PyTorch, Flask, Pydantic, PySpark, FAISS, ChromaDB, Matplotlib
Software & Tools: Git, Docker, Kubernetes, Amazon Web Services (AWS), Azure, Google Cloud Platform (GCP), Hadoop, Snowflake, Apache Spark, Scala, Tableau, Excel, Jira, Kafka, Confluence, CI/CD, Grafana, Jenkins, SageMaker, DataBricks, Airflow, MLflow, Power BI, Clustering, TensorFlow, CUDA, Key Performance Indicators (KPI)
Courses: Applied Machine Learning, Deep Learning, Natural Language Processing, Advanced Statistics, Econometrics and Time Series Analysis, Big Data Technologies, Predictive Analytics, Data Warehousing, Computer Vision, Cloud Computing, DevOps