Sumit De

Data Scientist

214-607-7911 | sxd0218@mavs.uta.edu| Arlington, TX | LinkedIn | GitHub

SUMMARY

Data Scientist with 3 years of experience applying advanced analytics and machine learning to influence product strategies and business outcomes. Skilled in SQL, Python, and cloud technologies to develop scalable solutions, define and monitor product metrics, and deliver actionable insights. Adept at cross-functional collaboration, data storytelling, and experimentation to drive data-informed decision-making.

SKILLS

- **Languages and IDEs:** Python, SQL, VS Code, Jupyter Notebook, PyCharm, Google Colab, JIRA
- **Frameworks & Libraries**: PyTorch, Langchain, TensorFlow, Scikit-learn, Keras, NumPy, Pandas, SciPy, Hugging Face
- **Cloud & Distributed Systems:** AWS (S3, Lambda, EMR, SageMaker, Bedrock), MLflow, Kubernetes, FAISS.
- **Software Engineering:** Version control (Git, Gitlab), Docker, CI/CD pipelines, REST API.
- **Machine Learning:** Supervised & Unsupervised Learning, Experiment Design, A/B Testing, Clustering, Model Evaluation.
- **Deep Learning:** CNN, ANN, RNN, LSTM.

EXPERIENCE

Charles Schwab, TX | Data Scientist *Jan 2024 – Present*

- Built scalable preprocessing pipelines to address class imbalance in internal operational risk datasets using techniques like SMOTE and feature normalization, improving data quality and enabling accurate detection of rare high-impact risk events.
- Developed and optimized XGBoost models to predict escalation likelihood and client churn within internal support systems, improving AUC-ROC by 20% and leveraging SHAP for actionable insight delivery to business stakeholders.
- Optimized investment portfolios using Monte Carlo simulations, resulting in a 2.5% increase in annualized returns while managing risk exposure. Boosted sales forecasting accuracy by 15% using ARIMA models, providing revenue growth trends.

Eaton Corporation, PA | Data Science Intern *May 2023 – Aug 2023*

- Developed an SVM-based cavitation detection model for centrifugal pumps, achieving 85.98% precision via advanced feature engineering, cross-validation, and hyperparameter tuning.
- Ingested and processed 134.27 million data points from 128 files using Python, downsampled the data by 97% for efficient analysis, and applied feature engineering for the selection of relevant features.
- Built scalable machine learning workflows using AWS (S3, Lambda, SageMaker), enabling parallel data processing and efficient model training in distributed cloud environments, accelerating deployment cycles.

Tata Consultancy Services (TCS), Bengaluru, India| Software Engineer *Apr 2021 - Jul 2022*

- Developed Python-based inter-node communication scripts for 5G core systems, handling HTTP requests and JSON responses in compliance with 3GPP standards, enabling seamless data flow between network components.
- Optimized integration of communication features by collaborating with frontend, QA, and engineering teams; conducted regression tests and leveraged A/B testing to select high-performing frontend versions.
- Utilized GitLab for version control and continuous integration, ensuring efficient code synchronization, collaboration, and high-quality software releases across cross-functional teams.

Zensar Technologies, Bengaluru, India| Data Scientist *Sep 2019 - Apr 2021*

- Implemented advanced machine learning based recommender systems, leveraging customer purchase history and browsing behavior to suggest personalized product recommendations, resulting in a notable 7% increase in conversion rate.
- Engineered customer segmentation models (K-Means, Decision Tree, Random Forest, and XGBoost) to identify distinct customer groups, enabling targeted marketing and increasing customer lifetime value by 9% on average.

DATA SCIENCE PROJECTS

Real-Time Attendance System using Face Recognition, ***Coursework Project***

- Developed prototype model of attendance system using face recognition. Made use of **deep learning (CNN)** and computer vision for face detection and **SVM** for face classification.
- Implemented triplet loss in a face recognition system to ensure embeddings of the same person are closer than those of different people, enhancing identification accuracy.
- Achieved accuracy of **97.31%** with dataset of **2100** training images and **900** test images.

Airport Analyzer, ***Business Symposium***

- Developed a **production-ready** web scraping and analysis pipeline using Selenium, NLP, and Langhian for optimized text chunking; deployed LLM-based insights to support airport authority decision-making from user reviews.
- Built a Streamlit user interface and implemented a cost-efficient **RAG workflow**, reducing inference costs by 70%; evaluated model performance using **ROUGE metrics** for quality assurance.

EDUCATION

Master of Science in Information Systems **May 2024**
University of Texas at Arlington, TX CGPA: 3.6/4

Bachelors in Electronics and Communication **Aug 2020**
PES University, Bangalore, India CGPA: 3.2/4