DHEERAJ SINGH SHANKAR

Data Scientist/Data Analyst

dheerajsinghshankar780@gmail.com

732-365-2144

PROFESSIONAL SUMMARY

* Over 8 years of experience in delivering production-ready solutions using Machine Learning, Deep Learning, and Natural Language Processing (NLP).
* Skilled in Python, TensorFlow, PyTorch, Scikit-Learn, Pandas, SQL Scripting, and Cloud Platforms (AWS, Azure, GCP).
* Proficient in predictive analytics, statistical analysis, data visualization (Power BI, Tableau), and identifying business opportunities through A/B testing and Causal Inference.
* Collaborative mindset with hands-on experience in MLOps, Reinforcement Learning, and Generative AI development.
* Experienced in Data Mining with large datasets of Structured and Unstructured data, Data Acquisition, Data Validation, Predictive modeling, and Data Visualization.
* Strong knowledge in Statistical methodologies such as Hypothesis Testing, Principal Component Analysis (PCA), Sampling Distributions, and Time Series Analysis.
* Expert in Python for data preprocessing, modeling, and analysis using Pandas, NumPy, Scikit-learn, Seaborn, and Matplotlib.
* Knowledge of Information Extraction, NLP algorithms coupled with Deep Learning.
* Experience in building various machine learning models using algorithms such as Linear Regression, Gradient Descent, Support Vector Machines (SVM), Logistic Regression, KNN, Decision Tree, Ensembles such as Random Forrest and Gradient Boosting Trees.
* Experienced in using cloud services like Amazon Web Services (AWS), such as EC2, S3, to work with different virtual machines.
* Experience in tuning algorithms using methods such as Grid Search, Randomized Search, K-Fold Cross Validation, and Error Analysis.
* Experience in Unsupervised learning working on social network datasets using K-means Clustering and Dimension Reduction methods.
* Experience in Natural Language Processing (NLP) and Time Series Analysis and Forecasting using the ARIMA model in Python and R.
* Experience in building and publishing interactive reports and dashboards with design customizations based on the stakeholders' needs in Tableau.
* Experienced in developing and designing ETL packages and reporting solutions using Tableau.
* Knowledge and experience in Agile Environments such as Scrum and using project management tools like Jira/Confluence and version control tools such as GitHub/Git.

EDUCATION

MS in Data Analytics and Visualization

BS in Computer Science

TECHNICAL SKILLS

|  |  |
| --- | --- |
| Programming Languages & Frameworks | Python (Pandas, NumPy, Scikit-learn, TensorFlow, XGBoost, LangChain, Hugging Face), R, SQL (SQL Scripting), Spark, JavaScript |
| Machine Learning | Predictive Modeling, Classification, Regression, Clustering, Ensemble Methods, PyTorch |
| Deep Learning | CNNs, RNNs, Transformers, Reinforcement Learning |
| Generative AI | LLMs, Prompt Engineering, RAG, LangChain, Hugging Face |
| Natural Language Processing (NLP) | Text Mining, Sentiment Analysis, Conversational AI, Chatbot Development, LLM Fine-Tuning, Named Entity Recognition (NER) |
| MLOps & Model Deployment | LLMOps, Model Deployment, ML Pipelines, Model Monitoring, MLFlow, Docker, REST APIs, Git, Linux, Data Confidentiality, Secure Pipeline Design |
| Cloud Platforms | AWS (EC2, S3), Azure (Blob Storage, AKS, Cognitive Services), GCP |
| Big Data Technologies | Spark, PySpark, Snowflake, ChromaDB, |
| Data Visualization & Tools | Jupyter Notebook, Power BI, Tableau, VS Code |
| Experimental Design & Statistical Analysis | EDA, Cross-validation, Feature Engineering, Hyperparameter Tuning, A/B Testing, Time Series Forecasting, Causal Inference, Reinforcement Learning, Precision-based model evaluation, Go-to-Market Metrics |

PROFESSIONAL EXPERIENCE

Valley National Bank, Morristown, NJ Sep 2024 - Present

Data Scientist/Data Analyst

Responsibilities:

* Participated in the full data science lifecycle, including data mining, cleaning, feature engineering, modeling, validation, and visualization.
* Built linear regression models and visualized customer NPS survey data to analyze satisfaction drivers across banking service branches.
* Developed predictive models to estimate customer likelihood to recommend banking products and services.
* Performed customer segmentation based on demographics and service usage patterns, offering data-driven service improvement recommendations.
* Developed classification and regression models using Python libraries such as Scikit-learn, Pandas, NumPy, Seaborn, and Matplotlib.
* Applied machine learning techniques including decision trees, logistic regression, SVM, random forest, XGBoost, KNN, and ensemble methods.
* Performed text analytics and NLP using NLTK for sentiment analysis on customer feedback and social media data.
* Built transformation pipelines for preprocessing financial data using techniques like scaling, imputing, and feature selection.
* Created LLM-powered chatbots for internal knowledge retrieval using Prompt Engineering and RAG (Retrieval-Augmented Generation) pipelines.
* Implemented clustering algorithms (K-Means, DBSCAN, and Hierarchical) for customer segmentation and behavior-based marketing in the banking domain.
* Conducted model optimization through cross-validation to reduce overfitting and improve generalization.
* Deployed machine learning workflows on AWS cloud virtual machines to analyze large-scale datasets.
* Performed sentiment classification using Naive Bayes and SVM on textual banking customer interactions.
* Leveraged ensemble learning (Bagging, Boosting) to enhance model accuracy for financial risk predictions.
* Processed diverse data formats (JSON, XML) to extract and standardize inputs for analytics pipelines.
* Conducted delivery route classification and package failure prediction using logistic regression and random forest techniques.

Environment: Python (Scikit-learn, Pandas, NumPy, NLTK, Seaborn, Matplotlib), AWS, Tableau Server, JSON/XML, Jupyter, Git

Sun Pharma - Princeton, NJ April 2024 – Aug 2024

Data Scientist/Data Analyst

Responsibilities:

* Conducted data profiling and exploratory analysis across multiple stages of clinical performance to identify behavioral trends and risk indicators.
* Engineered robust features by integrating demographic attributes, performance metrics, and textual data into unified modeling datasets.
* Applied advanced data preprocessing techniques, including cleaning, scaling, and transformation using Pandas and NumPy to ensure modeling readiness.
* Designed and fine-tuned LLM-based applications using OpenAI and Hugging Face for summarization, semantic search, and domain-specific Q&A.
* Developed Retrieval-Augmented Generation (RAG) pipelines combining FAISS/Pinecone with Sentence-BERT and OpenAI embeddings for context-aware responses.
* Built predictive models using XGBoost, Logistic Regression, Decision Trees, and Deep Learning (TensorFlow/Keras) to estimate outcome likelihoods.
* Implemented L1/L2 regularization techniques to address model overfitting in classification and regression use cases.
* Performed hyperparameter tuning using GridSearchCV and K-Fold cross-validation to optimize predictive performance.
* Developed ensemble pipelines utilizing FeatureUnion and FunctionTransformer to integrate categorical and continuous data sources.
* Leveraged NLP techniques via NLTK and SciPy for extracting insights from medical documents and narrative inputs.
* Created interactive visualizations using Matplotlib to present performance metrics across demographic and treatment cohorts.
* Applied prompt engineering and chaining logic to optimize LLM response accuracy and minimize hallucinations in clinical NLP workflows.
* Deployed multiple ML/DL models and conducted iterative testing in Signal Hub to identify the most effective solution.
* Used XGBClassifier and XGBRegressor in a hybrid modeling framework tailored for both categorical and continuous outcome prediction.
* Integrated GenAI-powered assistants into healthcare enterprise workflows using LangChain and APIs for contextual decision support and knowledge retrieval.

Environment: Python (Pandas, NumPy, SciPy, Scikit-learn, TensorFlow, Keras, NLTK), FAISS, Pinecone, OpenAI, Hugging Face, LangChain, Matplotlib, Signal Hub, XGBoost, GridSearchCV, APIs

Mphasis, Bengaluru, India Jan 2022 - Dec 2023

Data Scientist/Data Analyst

Responsibilities:

* Developed NLP-based virtual assistants using Kore.ai, JavaScript, and Python to automate customer support and streamline policy-related queries in the insurance and banking sectors.
* Designed and implemented Named Entity Recognition (NER) and intent classification models to boost chatbot resolution accuracy by 18%, reducing manual intervention.
* Integrated custom NLP pipelines for document parsing, enabling automated extraction of customer and transaction data from policy forms and loan applications.
* Built interactive, ML-powered customer analytics dashboards using Power BI, SQL, and Pandas, leading to a 25% increase in product feature adoption by business stakeholders.
* Leveraged LLMs for summarizing unstructured customer feedback and support tickets to enhance decision-making in claims processing and fraud detection.
* Designed scalable RESTful APIs to expose model predictions and recommendation systems to enterprise front-end systems.
* Worked with Azure Machine Learning for model deployment, versioning, and lifecycle management across development and production environments.
* Contributed to LLMOps pipelines to operationalize Large Language Models (LLMs) for use cases such as intelligent underwriting and customer Q&A.
* Built and tested microservices for GenAI components, including embedding generation, vector storage, and question-answering systems integrated with banking knowledge bases.
* Collaborated with DevOps teams to containerize ML services using Docker and deploy them on Azure Kubernetes clusters for auto-scaling and high availability.
* Conducted A/B testing and performance evaluation of GenAI-powered workflows, ensuring model precision and minimizing false positives in sensitive domains like credit risk.
* Automated real-time data ingestion and transformation pipelines for chatbot logs using Azure Data Factory and PySpark.
* Partnered with product and compliance teams to ensure responsible AI usage by validating outputs for bias and interpretability.
* Supported continuous monitoring of deployed models using custom logging and alerting systems integrated with Azure Monitor and Log Analytics.

Environment: Kore.ai, JavaScript, Python, Power BI, SQL, Azure Machine Learning, Azure Kubernetes Service (AKS), Azure Data Factory, PySpark, Docker, RESTful APIs, LLMOps, Pandas, Azure Monitor, Log Analytics

Brillio, Bengaluru, India May 2021 - Jan 2022

Data Scientist/Data Analyst

Responsibilities:

* Migrated telecom support portal from ExtJS to ReactJS, boosting page load performance by 40% and aligning with modern frontend frameworks.
* Integrated SpaCy NLP models to automate ticket classification; achieved a 20% improvement in routing efficiency.
* Applied CI/CD practices with Docker and Git to boost test coverage and streamline Agile Sprint Delivery.
* Conducted performance benchmarking before and after migration to quantify efficiency gains and align metrics with SLAs.
* Developed and deployed custom NLP pipelines using SpaCy to automatically classify and tag customer support tickets by issue type and severity.
* Improved ticket routing accuracy by 20% using machine learning classification models trained on historical labeled data.
* Engineered a data pipeline to collect and preprocess telecom support logs, applying text normalization, lemmatization, and named entity recognition.
* Integrated classification results with the internal support dashboard to enable real-time case triaging and escalation handling.
* Created visualization dashboards to track model predictions, routing times, and resolution performance using Power BI and Matplotlib.
* Conducted error analysis and confusion matrix reviews to iterate on SpaCy model performance and fine-tune thresholds.
* Applied CI/CD best practices by configuring Git hooks and automated testing to ensure model and UI integration consistency.
* Implemented unit and regression tests for Python NLP modules to improve reliability and accelerate feature deployment.
* Partnered with customer experience teams to identify high-impact use cases for ML-based automation in customer support workflows.
* Supported A/B testing of model-driven routing vs. manual routing to quantify impact on agent efficiency and customer satisfaction.

Environment: ReactJS, SpaCy, Python, Power BI, Docker, Git, CI/CD, Matplotlib, NLP, A/B Testing, Agile, Linux, REST APIs, PostgreSQL, VS Code, Jira

Mverve, Bengaluru, India Mar 2018 - Apr 2021

Data Scientist/Data Analyst

Responsibilities:

* Led hybrid app development using Apache Cordova, SQLite, and JavaScript, reducing deployment bottlenecks and improving platform reliability.
* Designed patient-provider recommendation logic simulating ML-based decision rules using hospital-provided data; supported backend APIs in Python and Node.js.
* Created internal KPI dashboards using Power BI and log analytics to visualize user flows and prioritize roadmap features.
* Simulated Machine Learning-based recommendation logic for patient-provider matching by analyzing structured hospital datasets and defining rule-based classification systems.
* Worked closely with clinical experts to fine-tune logic based on care pathways, specialty requirements, and patient history, enhancing recommendation accuracy and personalization.
* Designed data models to quantify user engagement, retention, and satisfaction across key customer lifecycle stages for both healthcare providers and patients.
* Built real-time monitoring dashboards for internal teams to prioritize roadmap features based on usage trends, issue frequency, and conversion metrics.
* Conducted data wrangling and preprocessing on structured and semi-structured healthcare datasets to ensure data quality and readiness for downstream analytics.
* Automated extraction and transformation of logs from mobile apps using custom Python scripts to feed into analytics pipelines.
* Utilized cohort analysis and segmentation techniques to personalize user communication strategies in CRM modules.
* Implemented A/B testing strategies to evaluate the effectiveness of different UI flows and recommendation rules, contributing to a 15% increase in appointment conversions.
* Supported customer relationship optimization by integrating user feedback loops into the data pipeline for continuous learning and product improvement.
* Applied statistical methods such as correlation analysis and logistic regression to validate hypotheses and drive strategic decisions.

Environment: Apache Cordova, JavaScript, Python, Node.js, SQLite, Power BI, SQL, Log Analytics, A/B Testing, Cohort Analysis, Logistic Regression, CRM

Subex, Bengaluru, India Mar 2017 – Feb 2018

Data Scientist

Responsibilities:

* Assisted in collecting data from clients and performed basic ETL operations to standardize data formats.
* Wrote SQL queries to extract sample datasets with key fields from SQL Server databases for analysis.
* Worked on machine learning algorithms to predict customer enrollment.
* Implemented K-Fold Cross Validation to test model reliability and generalization performance.
* Built Tableau dashboards and visual stories to present model summaries and key metrics to stakeholders.
* Supported SQL-based data extraction and transformation processes.
* Supported the creation of Entity Relationship Diagrams (ERDs) using Postgres SQL (Pgadmin) to guide database design.
* Visualized sales trends and seasonal patterns using Python libraries like Matplotlib and Seaborn.
* Explored and built forecasting models, including Linear Regression using Scikit-learn.
* Assisted in writing and updating stored procedures and triggers using T-SQL to support reporting and data processing needs.

Environment: Python, SQL Server, MS Excel, Tableau, Scikit-learn, Stats models, T-SQL, Matplotlib, Seaborn