Nageswara Rao Panja
Senior Data Scientist (Gen AI/ML)
Mobile no: +1 (940)758-2662
Email: panjanageswararao59@gmail.com
LinkedIn- https://www.linkedin.com/in/p-nageswara-rao-0521331b1

PROFESSIONAL SUMMARY:

- Overall, **7+ years of technical IT experience** in all phases of Software Development Life Cycle (SDLC) with skills in data analysis, design, development, testing and deployment of software systems.
- Experienced Senior Data Scientist seeking a challenging role, leveraging expertise in LLMs, AI product development, Chat/IVR systems, Speech to Text technologies, Vector Database & Graph DB, Transformer/Neural Network models, Open AI APIs, Langchain, Haystack frameworks, and advanced AI techniques.
- Experience developing data pipelines using AWS services including EC2, S3, Redshift, Glue, Lambda functions, Step functions, CloudWatch, SNS, DynamoDB, SQS.
- Proficiency in multiple databases like MongoDB, MySQL.
- Created **Snowflake Schemas** by normalizing the dimension tables as appropriate and creating a Sub Dimension named Demographic as a subset to the Customer Dimension. **Build multiple Data Lakes.**
- Experienced in Pivotal Cloud Foundry (PCF) on Azure VM's to manage the containers created by PCF.
- Extensive experience in Text Analytics, generating data visualizations using Python and creating dashboards using tools like Tableau, PowerBI.
- Provided full life cycle support to logical/physical database design, schema management and deployment. Adept at database deployment phase with strict configuration management and controlled coordination with different teams.
- Expertise coding in Python to manipulate data for data loads/extracts, statistical analysis/modeling & data munging.
- Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy. Experience in working on creating and running docker images with multiple microservices.
- Strong experience in Microsoft Azure Machine Learning Studio for data import, export, data preparation, exploratory data analysis, summary statistics, feature engineering, Machine learning model development and machine learning model deployment into Server system.
- Expertise in transforming business resources and requirements into manageable data formats and analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data.
- Worked with various text analytics libraries like Word2Vec, GloVe, LDA and experienced with Hyper Parameter Tuning techniques like Grid Search, Random Search, model performance tuning using Ensembles and Deep Learning.
- Skilled in System Analysis, E-R/Dimensional Data Modeling, Database Design and implementing RDBMS features.
- Well experience in Normalization and Denormalization techniques for optimum performance in relational and dimensional database environments.
- Experience in developing customized UDF's in Python to extend Hive and Pig Latin functionality.
- Experienced in building Automation Regressing Scripts for validation of ETL process between multiple databases like Oracle, SQL Server and MongoDB using Python.
- Proficiency in SQL across several dialects (MySQL, PostgreSQL, Redshift, SQL Server, and Oracle)

- Excellent communication skills. Successfully working in a fast-paced multitasking environment both independently and in a collaborative team, a self-motivated enthusiastic learner.

Technical Skills:

Big Data Ecosystem	Kafka, Flume, Cassandra, Amazon Web Services (AWS), EMR
Machine Learning Classification Algorithms	Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Gradient Boosting Classifier, Extreme Gradient Boosting Classifier,
	Support Vector Machine (SVM), Artificial Neural Networks (ANN), Naïve Bayes Classifier, Extra Trees Classifier, Stochastic Gradient Descent, etc.
Cloud Technologies	AWS, Azure, Google cloud platform (GCP)
IDE's	IntelliJ, Eclipse, Spyder, Jupyter
Ensemble and stacking	Averaged Ensembles, Weighted Averaging, Base Learning, Meta Learning, Majority Voting, Stacked Ensemble, AutoML – Scikit-Learn, MLjar, etc.
Databases	MySQL, DB2, MS-SQL Server.
Programming / Query Languages	C/C++, Java, SQL, Python Programming (Pandas, NumPy, SciPy, Scikit-Learn, Seaborn, Matplotlib, NLTK), NoSQL, PySpark, PySpark SQL, SAS, R Programming (Caret, Glmnet, XGBoost, rpart, ggplot2, sqldf), RStudio, PL/SQL, Linux shell scripts, Scala.
Data Engineer/Big Data Tools / Cloud / Visualization / Other Tools	Databricks, Hadoop Distributed File System (HDFS), Spring Boot, Flume, AWS, Azure Databricks, Azure Data Explorer, GCP, Google Shell, Linux, Big Query, Bash Shell, Unix, Tableau, Power BI, SAS, Scala AI TechnologiesLarge Language Models (LLMs), Transformer/Neural Networks, Open AI Whisper, Google TTS, Vector Database, Graph DB (Pinecone, Milvus, Neo4j) Frameworks: Langchain, Haystack Advanced AI Techniques: HyDE, MMR, LLM Reranking Tools: Git, Jupyter Notebook, Docker

WORK EXPERIENCE:

OPTUM, USA. **December 2023 – Present.**
Senior Data Scientist (Gen AI)

Responsibilities:

- Experienced Senior Data Scientist with expertise in cloud platforms such as Azure and AWS, data orchestration tools like Airflow, data warehousing solutions including Snowflake, and big data processing with Databricks. Proficient in programming languages including Python, Java, Node.js, and Spring Boot, with strong Unix skills.

- Led development projects utilizing Large Language Models (LLMs) for product enhancement, resulting in improved customer engagement and satisfaction.
- Collaborated with cross-functional teams to integrate LLMs into Chat and IVR systems, enhancing conversational AI capabilities and user experience.
- Implemented Speech to Text solutions using Whisper and Google TTS, achieving high accuracy rates and seamless voice interaction functionalities.
- Managed Vector Database & Graph DB solutions (Pinecone, Milvus, Neo4j) for data storage and retrieval, optimizing query performance and data insights.
- Developed and fine-tuned Transformer/Neural Network models for natural language processing tasks, such as sentiment analysis and entity recognition.
- Utilized Open AI APIs and Open Source LLMs (Llama2, Mistral, Mixtral) to build scalable and adaptable AI solutions for diverse business needs.
- Applied Langchain and Haystack frameworks for AI workflow orchestration, streamlining development processes and improving model deployment efficiency. Created Embeddings using MPNET and Ada models for semantic similarity and recommendation systems, enhancing content understanding and user personalization.
- Leveraged advanced AI techniques like HyDE, MMR, and LLM reranking to enhance semantic search capabilities, improving search result relevance and accuracy.
- Developed and maintained ETL/ELT solutions using Airflow to automate data workflows and improve operational efficiency.
- Designed and optimized data models in Snowflake for scalable and performant data storage and retrieval.
- Collaborated with cross-functional teams to ensure compliance with data standards and regulatory requirements.
- Mentored junior team members in best practices for data engineering and software development.
- Built data pipelines and workflows to ingest, transform, and load data from various sources into data warehouses.
- Implemented real-time data processing solutions using technologies like Kafka and Spark Streaming.
- Worked on performance tuning and optimization of SQL queries for efficient data retrieval.
- Automated data quality checks and monitoring processes to ensure data accuracy and reliability.
- Implemented data governance policies and procedures to ensure data integrity and security.
- Designed and implemented data streaming architectures using technologies like Kafka and Apache Flink.
- Conducted performance testing and capacity planning to optimize data processing and storage resources.
- Integrated machine learning models into data pipelines for predictive analytics and actionable insights.

- Collaborated with data scientists to deploy and operationalize machine learning models in production environments.
- Conducted code reviews and implemented best practices for software development and data engineering.
- Participated in agile development methodologies and contributed to sprint planning and retrospectives.
- Provided technical support and troubleshooting for production data pipelines and systems.
- Evaluated and recommended new technologies and tools to enhance data engineering processes and capabilities.
- Contributed to the development of data governance frameworks and data architecture standards.

Crowd Shakti, India **March 2018 – January 2022**
Data Scientist

Responsibilities:

- The planning and execution of data pipeline, data warehouse, and data lake architecture. This entails selecting data storage solutions and technologies that meet the needs of the business.
- creating and putting into use data models that enable effective and efficient data retrieval and manipulation. You will be expected to thoroughly comprehend and use database design ideas in your work.
- Integrating data from various sources into a single, cohesive system. This involves extracting, transforming, and loading data from different sources, such as databases, APIs, and file systems.
- Maintaining the data's accuracy and security. This entails creating and putting into practice data governance policies, data cleansing methods, and data validation processes.
- Leading a team of data engineers. This involves mentoring and guiding junior team members, overseeing project timelines, and ensuring that the team is meeting project goals and objectives
- Gathered business requirements, definition and design of the data sourcing, worked with the data warehouse architect on the development of logical data models.
- Created sophisticated visualizations, calculated columns and custom expressions and developed Map Chart, Cross table, Bar chart, Tree map and complex reports which involves Property Controls, Custom Expressions.
- Automated Diagnosis of Blood Loss during Emergencies and developed Machine Learning algorithm to diagnose blood loss.
- Created several types of data visualizations using Python and Tableau. Extracted Mega Data from AWS using SQL Queries to create reports.
- Performed reverse engineering using Erwin to redefine entities, attributes, and relationships existing database.

- Performed Regression testing for Golden Test Cases from State (end to end test cases) and automated the process using python scripts.
- Developed Spark jobs using Scala for faster real-time analytics and used Spark SQL for querying
- Generated graphs and reports using the gplot package in RStudio for analytical models. Developed and implemented R and Shiny application which showcases machine learning for business forecasting.
- Developed predictive models using Decision Tree, Random Forest, and Naïve Bayes.
- Used pandas, NumPy, seaborne, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms. Expertise in R, Matlab, python and respective libraries.
- Research on Reinforcement Learning and control (TensorFlow, Torch), and machine learning model (Scikit-learn).
- Hands on experience in implementing Naive Bayes and skilled in Random Forests, Decision Trees, Linear, and Logistic Regression, SVM, Clustering, Principal Component Analysis.
- Performed K-means clustering, Regression and Decision Trees in R. Worked on data cleaning and reshaping, generated segmented subsets using NumPy and Pandas in Python.
- Implemented various statistical techniques to manipulate the data like missing data imputation, principal component analysis and sampling.
- Worked on R packages to interface with Caffe Deep Learning Framework. Perform validation on machine learning output.
- Worked with Market Mix Modeling to strategize the advertisement investments to better balance the ROI on advertisements.
- Used Grid Search to evaluate the best hyper-parameters for my model and K-fold cross validation technique to train my model for best results.
- Worked with Customer Churn Models including Random Forest regression, lasso regression along with pre-processing of the data.
- Used Python 3.X (NumPy, SciPy, pandas, scikit-learn, seaborn) and Spark 2.0 (PySpark, MLlib) to develop variety of models and algorithms for analytic purposes.
- Performed Data Cleaning, features scaling, features engineering using pandas and NumPy packages in python and build models using deep learning frameworks
- Implemented application of various machine learning algorithms and statistical modeling like Decision Tree, Text Analytics, Sentiment Analysis, Naive Bayes, Logistic Regression and Linear Regression using Python to determine the accuracy rate of each model

- Implemented Univariate, Bivariate, and Multivariate Analysis on the cleaned data for getting actionable insights on the 500-product sales data by using visualization techniques in Matplotlib, Seaborn, Bokeh, and created reports in Power BI.

- Developed and maintained full-stack web applications using [programming languages and frameworks].

- Collaborated with cross-functional teams to design, implement, and deploy scalable and efficient cloud infrastructure on AWS.

- Utilized Terraform to automate the provisioning and configuration of AWS resources, ensuring infrastructure as code principles.

- Designed and implemented CI/CD pipelines for seamless deployment and continuous integration using [relevant tools].

- Worked closely with stakeholders to gather requirements and provide technical solutions to enhance application performance and user experience.

- Leveraged CloudFormation to deploy and manage AWS resources, ensuring secure and compliant infrastructure.

- **Environment:** Python, R, Scala, Scala AI Technologies Large Language Models (LLMs), Transformer/Neural Networks, Open AI Whisper, Google TTS, Vector Database, Graph DB (Pinecone, Milvus, Neo4j) Frameworks: Langchain, Haystack Advanced AI Techniques: HyDE, MMR, LLM Reranking ,Git, Jupyter Notebook, Docker, SQL, AWS, Azure, GCP, MySQL, PostgreSQL, Amazon Redshift, Google Big Query, Hadoop, Amazon S3, Apache Spark, Apache Kafka, and Apache Airflow, Tableau, Power BI, TensorFlow, PyTorch, Scikit-learn.

Crowd Shakti, India **November 2017 – March 2018**
Data Science

Responsibilities:

- Applied statistical techniques and models to analyze and understand time-dependent data, identify patterns, and extract meaningful insights.

- Build and implement time series forecasting models, such as ARIMA, SARIMA, exponential smoothing, or machine learning-based models like LSTM or Prophet, to predict future trends and outcomes.

- Cleanse and preprocess time series data, handle missing values, perform feature engineering to create relevant features for forecasting models.

- Evaluated the performance of different time series forecasting models using appropriate metrics and select the most accurate and appropriate model for the specific problem.

- Optimize model hyperparameters and parameters to improve forecast accuracy and generalization.

- Identified and handled outliers, anomalies, and seasonality in time series data to improve the accuracy and reliability of forecasting models.

- Conduct rigorous testing and validation of time series forecasting models using appropriate train-test splits, crossvalidation, and backtesting techniques.

- Monitor the performance of forecasting models over time, identify model degradation or drift, and update models as needed to ensure ongoing accuracy.
- Present and communicate forecasting results and insights to stakeholders effectively, using visualizations, reports, and presentations.
- Work closely with cross-functional teams, domain experts, and business stakeholders to understand their forecasting needs, align models with business goals, and drive data-driven decision-making.
- **Environment:** Python, Pandas, numpy, statsmodels, scikit-learn, TensorFlow, Keras, matplotlib, seaborn, Tableau, Power BI, Git, Apache Spark, Hadoop, Microsoft Azure, DynamoDB, Kibana, NOSQL, MYSQL

Education Details:

Masters in data science at university of North Texas. January 2022 - May 2023