**SPANDANA K**

**Data Scientist**

+1 3856854350 [kspandana011@gmail.com](mailto:kspandana011@gmail.com)

**PROFESSIONAL SUMMARY**

* 9**+ years** of experience in developing and implementing cutting-edge **and Machine Learning, Generative AI (GAI) Models and Algorithms** that scale across a massive volume of structured and unstructured data in **Manufacturing, Banking, Insurance and E-commerce** industries.
* Experienced in facilitating the entire lifecycle of a project Data Cleaning, Data Extraction, Data Pre-Processing, Data Transformation, Feature Engineering, Dimensionality Reduction, Model Building, Model Training and Hyper parameter Optimization, Model Deployment and Monitoring with large data sets of structured and unstructured data.
* Proficient in Machine Learning algorithm and Predictive Modeling including Regression Models, Decision Tree, Random Forests, Sentiment Analysis, Naïve Bayes Classifier, SVM, Ensemble Models, Boosting Methods.
* Proficient in Statistical Methodologies including Hypothetical Testing, ANOVA, Time Series, Principal Component Analysis, Factor Analysis, Cluster Analysis, Discriminant Analysis A/B Testing.
* Knowledge on time series analysis using AR, MA, ARIMA models.
* Expert in Deep Neural Networks, CNN, RNN, LSTM, GRU, Transformers, GAN, Auto Encoders, Attention based Models, Computer Vision.
* Has Experience in Customer Segmentation, Topic Extraction, Sentiment Analysis, Text Classification, Image Classification, Cluster Analysis, Outlier Detection, Anomaly Detection, Time Series Analysis, Recommendation systems, collaborative filtering, content-based filtering, matrix factorization
* Hands on experience in Large Language Models(LLMs), Generative AI models like Chat-GPT, BERT, RoBERTa, PaLM, Prompt engineering, Vector embeddings and the RAG framework.
* Experienced in designing, developing, and deploying GAI solutions.
* Proficient at a wide variety of Data Science programming languages **Python, R, SQL, PySpark, PyTorch.**
* Worked on serval python Libraries like NumPy, Pandas, NLTK, SciPy, Sci-Kit Learn (SK learn), TensorFlow, JAX, Keras, LangChain, Genism, SpaCy, TextBlob, Matplotlib, Beautiful Soup, Pickle, streamlit, gradio.
* Solid understanding in Dataiku DSS, effectively utilizing its data preparation, modeling, and collaboration features to streamline complex data workflows, develop predictive models, and drive data-informed decision-making
* Strong **SQL** programming skills, with experience in summarizing, transforming, segmenting, joining databases.
* Worked with **NoSQL** Database including **HBase**, **Cassandra** and **Mongo DB**.
* Proficient in data visualization tools like **Tableau** and **Power BI,** to create visually powerful, actionable and interactive reports and dashboards.
* Excellent Tableau Developer, expertise in building, publishing customized interactive reports and dashboards with
* Solid understanding of big data technologies like **Hadoop, Spark, HDFS, MapReduce, and Hive**.
* Experience in various **databases** such as MySQL, SQL, Oracle, NoSQL- MongoDB, Cassandra, HBase.
* Managed the machine learning lifecycle using MLflow, including model tracking, versioning, and experiment management.
* Proficient in all aspects of MLOps, from model development to training and deployment pipelines.
* Proficient in leveraging observability tools such as Splunk, Prometheus, and Grafana to analyze logs and metrics, enabling accurate diagnosis of system issues.
* Employed the ELK Stack (Elasticsearch, Logstash, Kibana) for log analysis and visualization, aiding in the monitoring of application and system logs.
* Proficient in OLAP data management using columnar databases like Amazon Redshift, Snowflake, and Databricks.
* Good at manage hosting plans for Azure Infrastructure, implementing and deploying workloads on Azure virtual machines.
* Hands on experience in **Azure ML Studio**, **GCP,** Google SQL,Big Query, Cloud Storage, G - cloud function, Cloud dataflow, Pub/Sub, GKE, AI Platform, Auto ML, Vertex AI and **AWS Sage Maker**.
* Strong expertise in container orchestration frameworks, particularly KubeFlow.
* Integrated ML workflows with C1/CD deployments for faster time to production.
* Experience in **Agile** and **Scrum** Methodologies
* Experience using version control systems **GIT** and **GitLab**.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Languages** | Python (NumPy, Pandas, Scikit-Learn, JAX, Matplotlib, Seaborn, Beautiful Soup, Prophet, pickle, LangChain, openai, streamlit), R, Scala, SQL |
| **Databases** | MySQL, MS SQL Server, MongoDB, Cassandra Teradata, snowflake, Google SQL |
| **Statistical Methods** | Hypothesis Testing, ANOVA, PCA, Time Series, Correlation (Chi-square test, t-test), Multivariate Analysis (ARIMA models), A/B Testing |
| **Data Visualization** | Tableau, Power BI, QlikView |
| **Machine Learning** | Linear Regression, Logistic Regression, Naïve Bayes, Decision Trees, Support Vector Machines(SVM), Clustering (K-means, K-NN), Random Forest, Gradient Boosting Trees, AdaBoost, XGBoost, Matrix Factorization, Principal Component Analysis (PCA), LDA, Natural Language Processing(NLP) |
| **Deep Learning** | LLMs, Artificial Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory (LSTM), Bi-LSTM, BERT, Transformers, Transfer Learning, TensorFlow, Keras, PyTorch |
| **Big Data** | Hadoop, Scala, Spark, MapReduce, Hive QL, HDFS, Sqoop, Pig |
| **Other Tools & Technologies** | Jupyter Notebooks, Anaconda, Databricks, Airflow, AWS ML, AWS SageMaker, GCP, Azure ML studio, NLTK, SpaCy, Genism, Kubernetes, Docker, Jenkins, Jira, Git, UNIX, |
| **Methodologies** | Agile, Scrum |

**EDUCATION**

* Bachelor of Technology in computer science from JNTUH, India

**PROFESSIONAL EXPERIENCE**

**Client: Toyota, US**

**Role: Data Scientist Timeline: Dec 2022 – till date**

**Responsibilities:**

* Designed and setup Enterprise Data Lake to provide support for various uses cases including **Analytics, processing, storing, and reporting of voluminous, rapidly changing data.**
* Involved in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP and coordinate task among the team.
* Developed and Demonstrated the POC, to migrate on-prem workload to Google Cloud Platform using **GCS, Google SQL** (**Big Query),** and **Cloud DataProc**.
* Utilized dimensional modeling techniques to structure data for OLAP purposes, ensuring granularity, integrity, and accessibility for comprehensive analysis.
* Aggregated and summarized data within columnar databases to create multidimensional cubes for efficient OLAP analysis.
* Created user defined views for accessing in **looker**.
* Analyzed data using SQL, R, Python, and presented analytical reports to management and technical teams.
* Employed Python to adeptly manage data conversion tasks to Transform and preprocess data to make it suitable for machine learning algorithms and Generative AI (GAI) models.
* Performed data cleaning and feature selection using Machine Learning package in **PySpark** and working with deep learning frameworks such as **TensorFlow, JAX, Keras** etc.
* Natural Language Processing (NLP) such as sentiment analysis, entity recognition, Topic Modeling and Text summarization was done using advanced python library such as NLTK, TextBlob, Spacy and Genism.
* Fine-tuned **large language models (LLMs)** to develop solutions in NLP and Generative AI.
* Optimized existing LLM training and serving, for improved model performance.
* Used Prompt engineering techniques to improve the performance of natural language processing (NLP) models, resulting in more coherent and contextually relevant outputs.
* Gathered, built, and annotated domain-specific datasets essential for training LLMs for a range of manufacturing tasks and applications.
* Analyzed model accuracy and bias, providing recommendations for corrective actions and improvements.
* Utilized parameter-efficient fine-tuning methods like p-tuning, LoRA to optimize the ML models.
* Implemented instruction tuning, RLHF, and parameter-efficient fine-tuning, and continually refined these processes through feedback and iterative improvement to fine-tune the LLM’s responses.
* Designed machine learning models tailored for recommendation systems. These models were designed to be scalable to handle large volumes of data.
* Implemented various techniques such as collaborative filtering, content-based filtering, and matrix factorization to enhance the recommendation engine’s performance.
* Delivered analysis support to recommendation systems by providing an online A/B test.
* Designed and implemented end-to-end systems for Data Analytics and Automation, integrating custom visualization tools using Python and Tableau.
* Wrote complex SQL statements to interact with the RDBMS database to filter the data and data analytics.
* Used Data Build Tool for transformations in ETL process, AWS lambda.
* Worked on scheduling all jobs using Airflow scripts using python. Adding different tasks to DAG’s and dependencies between the tasks.
* Involved in Creating, Debugging, Scheduling and Monitoring jobs using Airflow for ETL batch processing to load into Snowflake for analytical processes.
* Used AWS EMR to transform and move large amounts of data into and out of other AWS data stores and databases, such as Amazon Simple Storage Service (Amazon S3) and Amazon Athena.
* Managed the entire product site on Tableau and Qlik view while dealing with products relating to various clients.
* Conducted Data blending, Data preparation using SQL for **Tableau** consumption and publishing data sources to Tableau server.
* Implemented Machine learning models leveraging AutoML, Feature Store and Vertex AI.
* Customized machine learning models using libraries like TensorFlow and PyTorch within Vertex AI for fine-tuned model development.
* Used terraform to write Infrastructure as code and created Terraform scripts for EC2 instances, Elastic Load balancers and S3 buckets.
* Configured Azure Active Directory and managed users and groups
* Implemented AWS Step Functions to automate and orchestrate the **Amazon Sage Maker** related tasks such as publishing data to S3, training ML model and deploying it for prediction and Integrated Apache Airflow with AWS to monitor multi-stage ML workflows with the tasks running on Amazon Sage Maker.
* Create Athena data sources on S3 buckets for ad hoc querying and business dashboarding using Quick sight and Tableau reporting tools.
* Implemented MLOps practices, streamlining the model development lifecycle.
* Successfully deployed models to production using technologies like Docker, Kubernetes, and AWS SageMaker, ensuring low-latency and high availability.
* Monitored model performance in real-time, detecting and addressing drift issues through automated alerting and model retraining,
* Designed and deployed containerized applications using KubeFlow ensuring scalability and resource optimization.

**Environment:** LLM, NLP, GAI, Python, ELK, GCP, Big Query, Looker, Vertex AI, Amazon SageMaker, Airflow, Databricks, Datalake, Apache Spark, Snowflake, Tableau, CI/CD, Agile, A/B Testing, UNIX.

**Client: Goldman Sachs , US**

**Role: Data Scientist/ Data Analyst Timeline: Sept 2021 – Nov 2022**

**Responsibilities:**

* Participated in all phases of Machine Learning and Data Mining; data collection, data cleaning, developing models, validation, visualization.
* Used **Pandas, NumPy, Seaborn, SciPy, Matplotlib, SK-Learn, NLTK, Tensorflow, JAX, Keras**  in python for developing various machine learning  and deep learning algorithms.
* Implemented the end-to-end platform for performing user behavior analytics using unsupervised machine learning.
* Implemented Classification using supervised algorithms like Logistic Regression, Decision trees.
* Data transformation from various resources, data organization, features extraction from raw and stored.
* Identified outliers and inconsistencies in data by conducting exploratory data analysis (EDA) using python NumPy and Seaborn to see the insights of data and validate each feature.
* Expert in performing Text Mining and Text classification in **NLP** by using TfidfVectorizer.
* Developed NLP models for **Topic Extraction** and Sentiment Analysis.
* Performed price sensitivity and variation analysis across different marketing channels and conducted exploratory data analysis on variables such as lifetime value and profit score.
* Worked with a team of sales representatives to develop an MMM model to forecast sales.
* Utilized Market Mix Modeling (MMM) techniques to analyze and optimize marketing investments, effectively quantifying the impact of various marketing inputs on sales and Market Share, while ensuring a higher Return on Investment (ROI).
* Analyzed the data using various **machine learning algorithms** whether to extend/not credit limit to an existing applicant and to approve/not new credit line to a new applicant will likely result in profit or loss based on various circumstances like credit history, utilization rate, income, age, location, hard enquiries & number of deliquesces.
* Extracted terabytes of structured and unstructured data by using **SQL queries** and performed data mining tasks including handling missing data, data wrangling, feature scaling, and outlier analysis in python by importing pandas.
* Conducted data investigation, discovery & mapping tools to scan every single data record.
* Studied the feature distribution with the help of Probability Density Function, Cumulative Distribution Function, Percentiles, and Quintiles to draw some insights.
* Built decision tree model from the set of training data using the information entropy and the attribute with the highest normalized information gain is chosen to make the decision of credit approval.
* Validated models using cross-validation and loss function to measure model performance. Created **Confusion Matrix** and **ROC**.
* Built data pipelines, implemented code modularization involving package creation and co-developed **REST API’s** using Flask for production deployment.
* Built ML pipelines on Azure using best practices MLOps
* Visualized results in python using **Matplotlib**, **Seaborn** libraries and used Tableau to create the interactive dashboards to present results for team members, management, and clients.

**Environment:** Anaconda, Python, Jupyter, PyCharm, SSMS, Kafka, Tableau, Jira, Azure ML.

**Client: Teradata, India**

**Role: Data Scientist Timeline: Sept 2018 – July 2021**

**Responsibilities:**

* Gathered, analyzed, documented, and translated application requirements into data models and Supports standardization of documentation and the adoption of standards and practices related to data and applications.
* Participated in Data Acquisition with Data Engineer team to extract historical and real-time data by using **Sqoop**, **Pig, Flume, Hive, Map Reduce** and **HDFS**.
* Wrote user defined functions (UDFs) in Hive to manipulate strings, dates, and other data.
* Performed Data Cleaning, features scaling, features engineering using pandas and NumPy packages in **python**.
* Applied clustering algorithms i.e., Hierarchical, K-means using **Scikit** and **SciPy**.
* Performs complex pattern recognition of automotive time series data and forecast demand through the ARMA and ARIMA models and exponential smoothening for multivariate time series data.
* Delivered and communicated research results, recommendations, opportunities to the managerial and executive teams, and implemented the techniques for priority projects.
* Designed, developed, and maintained daily and monthly summary, trending, and benchmark reports repository in Tableau Desktop.
* Used GoogleSQL for all data transformations, leveraging a DAG orchestration tool to effectively manage GoogleSQL pipelines, ensuring streamlined and efficient data processing.
* Generated complex calculated fields and parameters, toggled and global filters, dynamic sets, groups, actions, custom color palettes, statistical analysis to meet business requirements.
* Implemented **visualizations** and views like combo charts, stacked bar charts, Pareto charts, donut charts, geographic maps, spark lines, crosstabs etc.
* Published workbooks and extracted data sources to **Tableau Server,** implemented row-level security and scheduled automatic extract refresh.

**Environment:** Machine learning, Linux, Python, R, Tableau, Hadoop, Map Reduce, HDFS, Hive, Pig, HBase, Sqoop, Flume, Oracle 11g, SQL Server.

**Client: Sapiens, India**

**Role: Data Analyst Timeline: May 2014 - Aug 2018 Responsibilities:**

* Worked with leadership teams to implement tracking and reporting of operations metrics across global programs.
* Worked with large data sets, automate data extraction, built monitoring/reporting dashboards and high-value, automated Business Intelligence solutions (data warehousing and visualization)
* Gathered Business Requirements, interacted with Users and SMEs to get a better understanding of the data.
* Performed Data entry, data auditing, creating data reports & monitoring all data for accuracy.
* Worked with OLAP data models aggregation, slicing, dicing, and drilling down into data for insights.
* Designed, developed, and modified various Reports.
* Performed data discovery and built a stream that automatically retrieves data from multitude of sources (**SQL** databases, external data such as social network data, user reviews) to generate KPI's using **Tableau**.
* Wrote ETL scripts in **Python, SQL** for extraction and validating the data.
* Create data models in Python to store data from various sources.
* Interpreting raw data using a variety of tools (**Python, R, Excel)**, algorithms, and statistical/econometric models (including regression techniques, decision trees, etc.) to capture the bigger picture of the business.
* Created and presented dashboards to provide analytical insights into data to the client.
* Translated requirement changes, analyzing, providing data driven insights into their impact on existing database structure as well as existing user data.
* Worked primarily on **SQL Server**, creating Store Procedures, Functions, Triggers, Indexes and Views using T-SQL.

**Environment:** SQL Server, ETL, SSIS, SSRS, Tableau, Excel, R, Python, Django