**Alekhya Mateti**

**+1(832)-400-2142**

[**Fazal@dataqinc.com**](mailto:Alekhyamateti07@gmail.com)

**LinkedIn ID:-** [**https://www.linkedin.com/in/amateti/**](https://www.linkedin.com/in/amateti/)

**Dallas, TX**

**Elevator:**

As a Senior Data Scientist with the Experience of around **9+ years**, I plan, create, and support Data Engineering programs that help firms make informed decisions. Machine learning, NLP, Data modeling, data analysis, and data visualization are other areas in which I am a specialist. These abilities help me convert raw data into insights that can be put to use. Additionally, I am familiar with agile approaches and have experience working in a dynamic team setting. I also have excellent verbal and written communication skills and problem-solving and analytical thinking abilities. I'm constantly studying and keeping up with the most recent developments in Machine Learning & Cloud Technologies.

**PROFESSIONAL SUMMARY:**

* Over 9 years of experience as a Data Scientist and Data Engineer handling multiple Data analytics and Data Engineering modeling techniques and strong cloud practices.
* Skilled in Python and Microsoft SQL Server, with 9 years of hands-on experience.
* Knowledge of NLP algorithms and Text Mining.
* Skilled in Advanced Regression Modeling, Correlation, Multivariate Analysis, Model Building, Business Intelligence tools, and application of Statistical Concepts.
* Developed predictive models using Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Cluster Analysis
* Experience in creating jobs for extract, transform and loading the data from source to destination databases.
* Excellent knowledge of Google Compute Engine, Google Cloud Storage, Google Cloud Load Balancing, and other services of the GCP family.
* Led a team of data scientists in developing and deploying machine learning models.
* Experienced in creating CI/CD pipelines on multiple integration tools, including SSIS, DataStage, and Azure data factory.
* Proficient in Snowflake's SQL-based querying language for data analysis and manipulation.
* Scheduled jobs for ETL tools and Proficient in creating reports and dashboards using PowerBi, Tableau, and SSRS.
* Experienced in building ML pipelines leveraging MLOps frameworks for model tracking, version controlling, monitoring, and deploying Machine Learning models into the production environment
* Played a crucial role in designing new architecture and creating a schema for a new database.
* Knowledgeable in agile sprint planning using Azure and Jira
* Experienced in setting up Snowflake data warehouses, configuring virtual warehouses, and managing resources effectively.
* Experienced in deploying code to Azure DevOps using pipelines.
* Familiar with the basic implications of different implementation decisions, such as distributed and parallel processing, through work on the IBM DataStage tool.
* Knowledgeable in big data technologies, including Spark, Hadoop, and Hive.
* Skilled in supervised and unsupervised machine learning algorithms.
* Completed a certification in Data Science from IBM on Coursera and Python programming language from Udemy.
* Hands-on experience in Python programming language.

**EDUCATION:**

* Bachelor of Technology– **Computer Science and Engineering**, JNTUH, India.
* Master of Science – **Data Sciences**, University of North Texas, College of Information**,** Denton, TX.

**PEER PARTICIPATIONS and CERTIFICATIONS:**

* Certified: **Python Programming.**
* Applied Data Science with **IBM**

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Languages | Python/R, SQL, Java/C#, Scala. |
| Methodologies/ Algorithms | Machine Learning /Deep Learning, NLP/NLU. |
| Databases | Microsoft SQL, MySQL, Postgres SQL, Oracle, MongoDB. |
| Big Data Ecosystems | Hadoop, HDFS, Hive, Spark, Kafka. |
| Data Visualizations | Tableau, Power BI, MATLAB. |
| Cloud Technologies | AWS, Sage Maker, GCP, Azure |
| Development Methodologies | Agile/Scrum, Waterfall. |
| Version Control Tools & Testing | Git, VM, GitHub. |
| Integrations | API Integration and Developments |

**PROFESSIONAL EXPERIENCE:**

**Client:** **Price Waterhouse Coopers PWC LTD Remote: Oct 2020 – Present**

**Role:** **Senior Data** **Scientist with NLP.**

* Extensively involved in all phases of data acquisition, data collection, data cleaning, model development, model validation and visualization to deliver data science solutions.
* Build recommendation engine models to solve the ticketing issues.
* Worked on NLP algorithms for documentation Sentiment analysis classification, text processing using NLTK, Text Blob, Spacy to find the information from the electronical files and text summarization.
* Set an Applied ML threshold at Dialog Flow that redirects the given utterances to the right chosen intent, which all performed in GCP.
* Proficiency in AI and machine learning frameworks, libraries and tools (e.g., LangChain, TensorFlow, Pinecone, PyTorch, GPT scikit-learn) including their pros and cons.
* Led a team of data scientists in developing and deploying machine learning models.
* Designed the Data processing architecture to enhance the Data visualizations and transform powerBI dashboards for business insights.
* Generated multiple dashboards on KPI’s and presented measurements and dimensions from powerBI and compared to tableau and provided data analysis.
* Identify and assess available machine learning and statistical analysis libraries (including regressors, classifiers, statistical tests, and clustering algorithms).
* Worked on Sentiment Analysis and Text Analytics to categorize comments into positive and negative clusters from different social networking sites.
* Experienced using Cloud Solutions- Google Vertex AI, CUDA CPU, GPU for training and deployment of models
* Successfully integrated various data sources into Snowflake, ensuring data consistency and accuracy.
* Applied NLP techniques to analyze customer feedback and reviews for the Snow Park.
* Deploy models using GCP Vertex AI, Amazon sagemaker, Airflow, Seldon, Flyte, ML flow, Kubeflow.
* Gathering business requirements from client and approach formulation and design methodology to match client requirements.
* Built a model end to end, from gathering requirements to deploying it on AWS using SageMaker.
* Built and tested different Ensemble Models such as Bootstrap aggregating, Bagged Decision Trees, and Random Forest, XGBoost, GBM to improve accuracy, reduce variance and bias, and improve stability of a model.
* Developed dashboards, user stories and reports using Tableau to analyze data associated with product performance, customer feedback and strategic decision making.
* Created ETL (Extract, Transform, Load) pipelines to automate data loading and transformation processes in Snowflake.
* Utilized Python libraries such as Pandas and NumPy to clean and preprocess large datasets, ensuring data quality and consistency.
* Utilized unsupervised learning techniques like k-means clustering to segment Snow Park visitors based on their preferences and behaviors.
* Performed data cleaning, including transforming variables and dealing with missing values and ensured data quality, consistency, and integrity using Pandas, NumPy
* Involved in training and deploying machine learning models to production using Vertex AI over GCP (Google Cloud Platform) and setting up model monitoring for document processing and understanding product.
* Utilized PCA and other feature engineering techniques to reduce the high dimensional data, applied feature scaling, and handled categorical attributes using one hot encoder of the sci-kit-learn library.
* Involved with the Bigdata developers’ team in developing Spark applications using Spark-SQL in Databricks on Azure for data extraction, transformation, and aggregation from multiple file formats for Analyzing& transforming the data to uncover insights into customer usage patterns.
* Improved query performance by optimizing SQL queries, utilizing Snowflake's automatic query optimization features, and indexing strategies.
* Developed and implemented machine learning models with Python libraries like Scikit-learn and TensorFlow, including linear regression, decision trees, random forests, and neural networks.
* Demonstrated an advanced command of Python, serving as the primary language for data manipulation, statistical analysis, and the development of machine learning models.
* Applied a deep understanding of Pandas for intricate data processing, extracting meaningful insights and facilitating data-driven decision-making.
* Utilized expert knowledge in machine learning, deploying TensorFlow and GPT PyTorch frameworks to craft highly accurate predictive models.
* Employed Spotfire and PowerBI to craft impactful visualizations and reports, translating raw data into actionable insights for stakeholders.
* Developed various machine learning models such as Logistic regression, KNN with Pandas, NumPy, Seaborn, Matplotlib, and Scikit-learn in Python.
* Implemented a Python-based distributed random forest via PySpark and MLlib
* Used cross-validation to test the model with different batches of data to find the best parameters for the model and optimize, which eventually boosted the performance.
* Created and maintained reports to display the status and performance of deployed model and algorithm with Tableau.

**Environment**: Python, pandas, TensorFlow, PyTorch, Spotfire, ML, DL, NLP, Tableau, MongoDB, GCP, PowerBI, Tableau, Azure and AWS.

**Client: Michaels Stores, NY April 2018 – Oct 2020**

**Role: Sr. Data Engineer / Data Scientist**

* Utilized domain knowledge and application portfolio knowledge to play a key role in defining the future state of large, business technology programs.
* Participated in all phases of data mining, data collection, data cleaning, developing models, validation, and visualization and performed Gap analysis.
* Conducted in-depth analysis of time-dependent data, identifying patterns and insights that informed strategic decision-making.
* Implemented Bayesian modeling principles to improve the accuracy and reliability of predictive models, leading to enhanced forecasting capabilities.
* Collaborated with cross-functional teams to integrate data science solutions into existing business processes, ensuring seamless implementation
* Implemented a Python-based distributed random forest via Python streaming.
* Created ecosystem models (e.g., conceptual, logical, physical, canonical) that are required for supporting services within the enterprise data architecture (conceptual data model for defining the major subject areas used, ecosystem logical model for defining standard business meaning for entities and fields, and an ecosystem canonical model for defining the standard messages and formats to be used in data integration services throughout the ecosystem).
* Used Pandas, NumPy, seaborn, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms and utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis.
* Spearheaded Chatbots development initiative to improve customer interaction with application.
* Developed the Chatbots using api.ai.
* Designed and maintained data models in Snowflake to support efficient querying and reporting.
* Created informative data visualizations using Python tools like Matplotlib, Seaborn, and Plotly, enabling stakeholders to grasp complex data patterns and trends.
* Designed and trained a time series forecasting model to predict weather conditions at the Snow Park.
* Familiarity with MLOps tooling from GCP's Vertex AI, enabling the efficient development, deployment, and monitoring of machine learning models in production environments
* Hosted on an AWS compute instance and using AWS services like SageMaker, Athena, Redshift, and S3.
* Automated CSV to chatbot-friendly JSON transformation by writing NLP scripts to minimize development time by 20%.
* Build Vertex AI pipelines. Use the Kubeflow Pipelines SDK to build scalable ML pipelines
* Conducted studies and rapid plots and used advanced data mining and statistical modeling techniques to build a solution that optimizes the quality and performance of data.
* Designed and trained deep learning models with Python frameworks such as Keras and PyTorch GPT for tasks like image classification, object detection, and sequence-to-sequence modeling.
* Analyzed large data sets, applied machine learning techniques, developed predictive and statistical models, and developed and enhanced statistical models by leveraging best-in-class modeling techniques.
* Managed Snowflake resources to handle varying workloads, ensuring optimal resource allocation and cost control.
* Developed Map Reduce/Spark Python modules for machine learning & predictive analytics in Hadoop on AWS.
* Developed machine learning models using AWS SageMaker, leveraging its built-in algorithms and frameworks like TensorFlow and PyTorch.
* Demonstrated exceptional proficiency in Python, utilizing it for end-to-end data science projects, data manipulation, and statistical analysis.
* Leveraged Pandas for advanced data processing, ensuring efficient extraction of valuable insights from complex datasets.
* Applied extensive knowledge of machine learning frameworks, including TensorFlow and PyTorch, to design and implement predictive models.
* Achieved improved accuracy and efficiency through the development of sophisticated machine learning algorithms.
* Collaborated on SAP data migration projects, including data cleansing, transformation, and loading (ETL) processes.
* Contributed to SAP best practices development and adherence, ensuring efficient and reliable system operations.
* Worked on customer segmentation using an unsupervised learning technique - clustering.
* Led data modeling initiatives to create efficient data structures and optimize data processing pipelines, ensuring data accuracy and performance.
* Managed Azure Data Lake Storage Gen2 (ADLS Gen2) for storing and processing vast volumes of structured and unstructured data, facilitating data-driven decision-making.
* Collaborated closely with data scientists and analysts, leveraging Databricks and Azure Synapse to develop data-driven insights, analytics, and machine learning models.
* Worked extensively with Azure cloud services to design and implement data solutions, harnessing the power of the cloud for scalability and cost-effectiveness.
* Utilized SQL for querying, transforming, and optimizing data for various business applications, ensuring data consistency and quality.
* Leveraged ERP systems, including SAP S/4HANA and JDE (JD Edwards), to integrate and manage critical business data, enabling streamlined operations.
* Managed product allocation processes, specifically in the Energy Components domain, to optimize resource distribution and support efficient production planning.
* Utilized Spark, Scala, Hadoop, HBase, Kafka, Spark Streaming, MLlib, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
* Designed and implemented system architecture for Amazon EC2 based cloud-hosted solution for the client.

**Environment:** Python, pandas, TensorFlow, PyTorch, Spotfire, SQL, Tableau, Cluster analysis, NLP, Spark, Mongo DB, logistic regression, Azure, HDFS, SVM, JSON, XML, AWS, Scala, PowerBi.

**Client:**  **Centene, MO Apr 2016 – Mar 2018**

**Role: Data Engineer / Data Scientist**

* Extensively involved in all phases of data acquisition, data collection, data cleaning, model development, model validation, and visualization to deliver data science solutions.
* Documented requirements through functional specifications and technical design documents.
* Created complex SQL queries and database objects to pull and manage data.
* Developed, tested, and deployed standard reports and dashboards based on business requirement document specifications.
* Updated tracking documents, generated different reports, and gave presentations on findings.
* Utilized SageMaker's data processing capabilities to clean, preprocess, and transform raw data into a format suitable for model training.
* Implemented data wrangling techniques in Python, including merging, pivoting, and reshaping data frames, to prepare data for analysis and modeling.
* Engaged with business users to gather, define, and document reporting requirements.
* Use pre-built components for interacting with Vertex AI services, provided through the Google Cloud pipeline components library Schedule a pipeline job with Cloud Scheduler. Vertex AI pipelines are similar to kubeflow pipelines.
* Performed iterated-based error analysis to monitor the production pipelines, data drift or concept drift issues, and monitoring the factors (server load, Input Null values) are causing model degradation.
* Engineered and selected relevant features from raw data using Python, improving model performance and interpretability.
* Performed data management techniques by combining the flexibility, cost-efficiency, and scale of databricks of data lakehouse with databricks on Azure and aligned ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data.
* Utilized advanced data processing techniques with Pandas to extract valuable insights from complex datasets, enhancing decision-making processes.
* Applied expertise in machine learning frameworks, including TensorFlow and PyTorch, to design and implement predictive models, resulting in improved accuracy and efficiency.
* Leveraged Spotfire and PowerBI for data visualization and reporting, transforming raw data into actionable insights for stakeholders.
* Conceived, developed, and implemented a wide range of BI solutions.

**Environment**: Python, pandas, TensorFlow, PyTorch, Spotfire, ML, DL, NLP, AWS Sagemaker Tableau, MongoDB, GCP, PowerBI, Tableau, Azure and AWS.

**Client:** **GGK Tech, Hyderabad, India. June 2014 – Dec 2015**

**Role: Data Engineer.**

* Worked with team members to promote excellent customer service and a pleasant work environment.
* Attended and participated in meetings and brainstorming sessions with team members.
* Supported various program activities, assisting with administrative tasks.
* Assisted in systematic research, investigations, and special projects to support the work of assigned departments.
* Collaborated with cross-functional teams to design, develop, and maintain a cloud-based enterprise data warehouse for the State of Georgia's Department of Community Health (DCH) as part of the EASE project.
* Played a key role in implementing Data Vault 2 methodologies and WhereScape Data Warehouse Automation to build and maintain the data models in Snowflake.
* Worked closely with business stakeholders to understand reporting requirements and translated those requirements into conceptual, logical, and physical data models for the cloud data warehouse.
* Created database scripts and power scripts, including data definition language, database schemas, and roles, for all layers of the cloud data warehouse, ensuring data integrity and quality.
* Developed, implemented, and automated testing scripts using Validator to ensure the integrity and quality of the data warehouse.
* Wrote and deployed code to all data warehouse environments, maintaining version control and ensuring smooth deployments.
* Designed and implemented role-based security throughout the data warehouse to safeguard sensitive information and control data access.
* Gather, analyze, and translate business requirements into functional specification documents and share for peer review.
* Leveraged SageMaker AutoML for automated model selection and hyperparameter tuning, streamlining the model development process.
* Leverage enterprise data science methods across various platforms including R, Python, SQL, Google Cloud Platform (BigQuery, Vertex AI), Jira, Confluence, and GitHub.
* Analyzed the customer data and business rules to maintain the data quality and integrity.
* Developed SQL queries to create tables, extract/merge to bring data together from various sources.
* Involved in the data cleaning procedures, data alignment, and data profiling.
* Maintained accurate records and files pertaining to departmental areas of responsibility.