**Vijay Muni Reddy M**

**+1 973-449-0119 |** [**vijayreddymuni@gmail.com**](file:///C:\Users\bhask\AppData\Local\Microsoft\Windows\INetCache\IE\GBU0DZEV\vijayreddymuni@gmail.com)

**Data Scientist**

**PROFESSIONAL SUMMARY**

* **5** years of professional experience in the Data Science field, specializing in statistical analysis, data mining, machine learning, and deep learning.
* Led a team to develop data-driven solutions for business problems, focusing on analytics and basic machine learning techniques for image analysis and object detection.
* Expertise in Python programming for developing machine learning models, automating data processing tasks, and optimizing performance in data analysis and model deployment.
* Experienced in developing and implementing machine learning models using Poarch and TensorFlow, focusing on data analysis, predictive modeling, and automation.
* Experienced in deploying machine learning models using AWS Sage Maker, processing large datasets with PySpark, and utilizing Hugging Face for model management and training.
* Skilled in optimizing machine learning models by fine-tuning and training on specific datasets to improve accuracy and performance.
* Experience in developing different statistical Machine Learning solutions, text summarizations, sentiment analysis, recommendation systems, forecasting models, fraud detection models, object detection, and segmentation to various business problems
* Experience in Machine Learning (ML), Predictive Modeling, Natural Language Processing (NLP), and Deep Learning algorithms Strong DevOps skills, including experience with CI/CD pipelines, Docker, Kubernetes, and Terraform, streamlining model deployment and lifecycle management.
* Knowledgeable in a wide range of Machine Learning algorithms, including Ensemble Methods (Random Forests), Linear and Logistic Regression, Support Vector Machines (SVM), Deep Neural Networks, Extreme Gradient Boosting, Decision Trees, K-Means, K-NN, Gaussian Mixture Models, Naive Bayes, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM)
* Experience in data visualization using Python and PowerBI software to design dashboards for presentation and publication
* Strong DevOps skills, including experience with CI/CD pipelines, Docker, Kubernetes, and Terraform, streamlining model deployment and lifecycle management.

**TECHNICAL SKILLS**

**Programming:** Python, Java, SQL, NumPy, Pandas, scikit-learn, TensorFlow, PyTorch, Matplotlib, Stream lit.

**Software Tools:** Git, GitHub, ETL, Postman, REST API.

**Cloud Platforms:** AWS, Azure.

**Data Engineering:** Spark, DBT, Data Modeling

**DevOps and CI/CD Tools**: Docker, Kubernetes, Jenkins, Terraform

**ML/DL:** Linear Regression, Logistic Regression, Decision tree, Random Forest, XGBoost, SVM, PCA, CNN, RNN, LSTM, GAN, Transformers, Generative AI

**Data visualizations:** Tableau, Power BI, Power Query, Power Automate.

**EXPERIENCE**

**Client: Baycare Health, New York City July 2023 – Present**

**Role: Data Scientist**

**Responsibilities:**

* Implemented a machine learning-based system for classifying intent-based user instructions in electric automation, enhancing robotic automation systems' efficiency and adaptability
* Worked on the multimodal, developing a visual question-answering system utilizing hugging face and Streamlit
* Drove the development of intricate crowd-counting models, collaborating across teams to seamlessly integrate cutting-edge computer vision solutions. Achieved a remarkable 95% accuracy rate, solidifying project success.
* Spearheaded a YouTube video summarization project using OpenAI’s ChatGPT, resulting in an engaging web app with Stream lit
* Designed the Machine Learning data pipeline for regular monitoring and performance evaluation of the deployed ML models
* Leveraged advanced machine learning techniques to develop predictive models and algorithms for financial data analysis, risk assessment, and fraud detection with Python programming and its associated libraries, i.e. pandas, NumPy, scikit-learn, TensorFlow
* Conducted data preprocessing, cleaning, and feature engineering to prepare datasets for modeling, ensuring data quality and integrity
* Designed and optimized end-to-end ML pipelines, implementing CI/CD practices to automate model training, testing, and deployment.
* Implemented models using TensorFlow and PyTorch, conducting hyperparameter tuning and model evaluation to achieve optimal performance across various environments.
* Applied advanced statistical analysis techniques, such as regression, time series analysis, and clustering, to extract meaningful insights and identify patterns in financial data
* Built and maintained data pipelines using Apache Spark and DBT, ensuring high-quality data flows for ML model training
* Owned the complete MLOps lifecycle, including data monitoring, code refactoring, and the development of robust model monitoring workflows for efficient model lifecycle management
* Applied strong knowledge of data structures, algorithms, and software engineering principles to create robust AI solutions
* Designed and optimized containerized environments for model deployment using Docker and Kubernetes (K8s), ensuring scalability and ease of maintenance.
* Collaborated with cross-functional teams to define AI project requirements and objectives, ensuring alignment with overall business goals
* Conducted ongoing research to stay updated with the latest advancements in generative AI, machine learning, and deep learning techniques, identifying opportunities for integration into products and services
* Developed RESTful APIs in Java to expose machine learning model predictions, enabling seamless integration with data analytics platforms.
* Designed and implemented ETL workflows using Java and Apache Spark, optimizing data transformations for analytics and reporting.

**Environment:** Python, TensorFlow, AWS SageMaker, Apache Spark, DBT, ETL workflows, Data Preprocessing, Feature Engineering, Docker, Kubernetes (K8s), MLOps, CI/CD, RESTful APIs, Streamlit, Hugging Face, Git, GitHub

**Client: Splunk, Hyderabad, India Aug 2021 – Sep 2022**

**Role: Python Developer**

**Responsibilities:**

* Participated in requirement gathering and collaborated with business teams to understand data needs for real-time analytics.
* Developed Python scripts for data ingestion into Splunk, ensuring smooth integration with various data sources like AWS S3 and AWS Kinesis.
* Assisted in building and evaluating Machine Learning (ML) models for predictive analytics, focusing on regression and classification tasks using scikit-learn and pandas.
* Worked on integrating AWS Sage Maker to automate model training and deployment processes for data analytics applications.
* Developed and deployed RESTful APIs using Flask to facilitate data access, model predictions, and integrate with the monitoring system.
* Assisted in writing Python scripts to preprocess data, clean datasets, and ensure better quality for ML model training.
* Applied basic data science techniques like feature engineering, data normalization, and hyperparameter tuning to improve model performance.
* Worked on AWS Glue for ETL processes, ensuring seamless data transformation and integration between various data sources and Splunk.
* Created Python-based data pipelines for streaming and batch processing using AWS Lambda, AWS SQS, and AWS Kinesis.
* Assisted in maintaining Splunk deployment and developed Python-based tools for system monitoring and alerting based on model predictions.
* Wrote Python scripts to automate routine tasks, reducing manual intervention and increasing operational efficiency.
* Collaborated with cross-functional teams (DevOps, QA, Data Science) to ensure the smooth integration of Python scripts in Splunk’s CI/CD pipeline.
* Troubleshot and fixed minor issues in Python-based applications, ensuring smooth performance in production.
* Assisted in the deployment of ML models using AWS Sage Maker and integrated predictions into Splunk dashboards for real-time insights.
* Used scikit-learn in modeling various classification, regression, and clustering algorithms including support vector machines, random forests, gradient boosting, k-means
* Deployed models using Docker and Kubernetes, enhancing deployment reliability, and enabling efficient scaling within the cloud environment.
* Conducted data preprocessing, cleaning, and feature engineering to prepare datasets for modeling, ensuring data quality and integrity
* Involved in developing and testing SQL scripts for report development, Tableau reports, dashboards, and effectively addressed performance issues

**Environment:** Python, Flask, Splunk SDK, AWS Lambda, AWS S3, AWS Kinesis, AWS Glue, AWS Sage Maker, AWS SQS, MySQL, scikit-learn, pandas, RESTful APIs, Jenkins, GitLab, Linux

**Client: PayU, India June 2019- Aug 2021**

**Role: Data Analyst**

**Responsibilities:**

* Developed 30+ visually compelling and interactive Power BI dashboards and reports, leveraging DAX (Data Analysis Expressions) and Power Query M for advanced data modeling and data visualization, providing key stakeholders with actionable insights.
* Responsible for performing data analysis, pre-processing and validation of large datasets loaded to the big data environment and providing feedback about data quality to the data load team before the data can be consumed by data scientists.
* Conducted data analysis using Power BI, Python, and SQL to support business operations and strategic initiatives. Created and maintained dashboards and reports for various stakeholders, enabling data-driven decision-making.
* Created and maintained Data Warehouse data model following set standards, utilizing both Relational and Dimensional Modeling techniques. Provide data quality advisory services to clients and to internal stakeholders. Recommend maintenance enhancements to data acquisition processes to improve accuracy of data warehouse
* Ingested data from multiple data sources using a combination of SQL, Google Analytics API, Salesforce API using Python to create data views to be used in BI tools like Tableau & Power BI
* Expertise in writing complex DAX functions in Power BI and Power Pivot. Created SSIS packages to integrate the data coming from flat files.
* Worked on Azure cloud in spinning of new resources under resource group. Worked on Azure Data factory to load the data from on premise to Azure Data Lake. Scheduled Automatic refresh and scheduling refresh in Power BI service.
* Profound knowledge of Software Development Life Cycle (SDLC) with a thorough understanding of various phases like Requirements Gathering, Analysis, Design, Development, Testing, and Agile methodologies.
* Performed complex data analysis and generated insights by applying statistical techniques, data mining algorithms, and visualization tools.
* Helped create and maintain dashboards and reports to monitor key performance indicators and business metrics.

**Environment:** Power BI, DAX, Power Query M, Python, SQL, Salesforce API, Tableau, SSIS, Azure Data Factory, Azure Data Lake, Power Pivot, Data Warehouse Modeling, Agile SDLC, Data Analysis, Data Preprocessing.

**EDUCATION**

* **New Jersey Institute of Technology**, Newark, NJ

Master of Science in Machine Learning (2023)

* **Nitte Meenakshi Institute of Technology**, Bangalore, India

Bachelor of Engineering in Computer Science (2022)

**Leadership/Certification**

AWS Certified Data Engineer – Associate, Microsoft Certified: Power BI Data Analyst Associate, Microsoft Certified Data Engineer – Associate, Complete SQL Bootcamp, Introduction to LLM, Azure Databricks for Data Engineers (PySpark, SQL).

**Key Achievements**

Developed predictive models and fraud detection algorithms for financial data, improving risk assessment accuracy.

Led the implementation of an automated data pipeline using AWS and Python, enhancing data processing efficiency.

Created and deployed 30+ Power BI dashboards, enabling data-driven decision-making and increasing business insights.

Ranked in the top 10% in the IBM Innovation Club contest, showcasing advanced problem-solving skills in software development and technical innovation

Successfully completed multiple projects during university-level hackathons, demonstrating strong collaboration, coding skills, and the ability to meet tight deadlines

**Notable Projects:**

#### **1. Brain Tumor Detection from MRI Data using CNN**

**Technologies:** Python, TensorFlow, Keras, CNN, ResNet, VGG16, Scikit-learn, AUC-ROC, Decision Trees, Random Forest

* Developed a **deep learning-based MRI analysis system** to detect brain tumors, leveraging **transfer learning with pre-trained models (VGG16, ResNet)** for enhanced accuracy.
* Addressed **class imbalance** and tumor variability by implementing **data augmentation and feature engineering techniques**, improving generalization and robustness.
* Designed a **CNN-based classification pipeline**, achieving **93% accuracy**, validated using **AUC-ROC, Sensitivity, Precision, and F1-Score** to ensure high performance.
* Automated **image preprocessing workflows**, including normalization, segmentation, and feature extraction, reducing data processing time by **40%**.

#### **2. Reddit ETL Pipeline using Apache Airflow, Docker, and AWS**

**Technologies:** Apache Airflow, Docker, Python, Reddit REST API, AWS S3, PostgreSQL, Apache Spark, PRAW, CI/CD, Terraform, Kubernetes

* Designed and implemented an **end-to-end ETL pipeline** to extract, process, and store top Reddit posts from the **Data Engineering subreddit**, enabling real-time insights.
* Developed **Airflow DAGs** to automate **data extraction, transformation, and storage**, ensuring **scalability, high availability, and automation**.
* Integrated **Reddit REST API (PRAW)** for secure data retrieval, leveraging **OAuth authentication** and data ingestion pipelines for structured processing.
* Configured **Apache Airflow with Docker & PostgreSQL**, optimizing data transformation workflows and **improving query performance by 35%**.
* Implemented **real-time monitoring and logging** using Airflow, enabling **error tracking and automated alerts** for pipeline health.
* Deployed the pipeline on **AWS S3 & PostgreSQL**, leveraging **Terraform and Kubernetes** for auto-scaling and optimized cloud resource management.

#### **3. MongoDB-Based Student Performance Prediction System**

**Technologies:** Python, MongoDB, Streamlit, Scikit-learn, Pickle, StandardScaler, Label Encoding, Power BI

* Developed a **student performance prediction system**, integrating **MongoDB for real-time data storage** and **Scikit-learn for ML-based predictions**.
* Built an **interactive Streamlit UI**, allowing students and educators to input study hours, sleep patterns, and past performance to generate insights.
* Engineered **a machine learning model (Logistic Regression)** for predictive analysis, improving academic decision-making and student engagement.
* Implemented **data preprocessing techniques** (StandardScaler, LabelEncoder) to ensure model consistency, boosting accuracy and reducing bias.
* Automated **real-time logging and data storage** in **MongoDB Atlas**, enabling historical trend analysis and improved student tracking.
* Designed **Power BI dashboards** from stored MongoDB data, visualizing **academic performance trends and key predictors** for actionable insights.