Bhavya Raj B

Sr. Data Scientist/ ML engineer

**SUMMARY**

* Data Scientist/MLops Engineer with 9 Years of Experience. Proficient in designing, deploying, and maintaining scalable machine learning infrastructure and operations in e-commerce, healthcare, finance, and technology industries.
* Experience in **productionizing Machine Learning pipelines** on Google Cloud Platform which performs data extraction, data cleaning, model training and updating the model on performance basis.
* Expertise in managing end-to-end data pipelines with **RDF** **Graphs**, **Apache** **Beam**, and **Apache** **Nifi** ensuring seamless data flow and integration.
* Extensive experience in building and maintaining production ML pipelines with GCP resources such as **BigQuery, Cloud Composer, Compute Engine, Kubernetes clusters, and GCP storage buckets.**
* **Proficient in containerization and orchestration** using Docker and Kubernetes for scalable deployment of ML applications.
* Implemented comprehensive **monitoring** and logging using **Prometheus, Grafana, ELK Stack, and Azure Monitor** to ensure model and infrastructure health.
* Designed and implemented **MLOps** solutions on AWS, GCP, and Azure, leveraging cloud-native services for scalable and resilient deployments.
* Utilized tools such as Apache **Airflow**, **Luigi**, and **Kubeflow** Pipelines to automate end-to-end ML workflows, from data ingestion to model deployment.
* Developed comprehensive monitoring and alerting systems using tools like **Prometheus**, **Grafana**, and **CloudWatch** to ensure high availability and performance.
* Implemented automated **model** **drift** **detection** **systems** to monitor and alert model performance degradation over time.
* Collaborative work with cross-functional teams and utilization of diverse technologies (**Python**, **Scala**, **TensorFlow**, **PyTorch**).
* Hands on solving problems which brings significant business value by building predictive & forcasting models utilizing **structured & unstructured data**.
* Hands - on experience in Machine Learning algorithms such as Linear Regression, GLM, CART, SVM, KNN, LDA/QDA, Naive Bayes, Random Forest, SVM, Boosting.
* Hands on experience in creating data visualizations, dashboards in a **Tableau** desktop.
* Experience in building data warehouses, data marts and data cubes for creating **PowerBI** reports to visualize various key performance indicators of business.
* Utilized python libraries namely **Pandas**, **matplotlib** and **plotly** for performing data analysis, data visualizations and predictions.
* Utilized python’s flask framework for building REST APIs on top of Data Lake (**BigQuery**, **Cloud** **SQL**).
* Achieved Continuous Integration &Continuous Deployment (**CI/CD**) for applications using Git, Azure Devops.
* Experience with **Test driven development** (TDD), **Agile** methodologies and **SCRUM** processes.
* Experience in version control and collaboration tools like **Git** and **source** **tree**.

**TECHNICAL SKILLS**

**Languages:**Python, R, SQL, Java Script

**ML/AI:**Tensorflow, Kera’s, Scikit-learn, Prophet, PySpark, NLTK, Airflow, Pandas, OpenCV

**Data Base:**MySQL, SQL server, PostgreSQL, MongoDB

**Reporting Tools:**Tableau, Power BI, Wavefront

**Predictive and Machine Learning:**Regression (Linear, Logistic, Bayesian, Polynomial, Ridge, Lasso), Classification (Logistic Reg., two/multiclass classification, Boosted Decision Tree, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machines, k-Nearest Neighbours, Neural Network, and various other models), Clustering (K-means, Hierarchical), Anomaly Detection, LSTM, RNN

**Cloud:**Google Cloud Platform, Pivotal Cloud Foundry, Azure, AWS

**Cloud Resources**: Azure Data bricks, AWS Glue, GCP BigQuery, Cloud Composer, Dataflow, Data Proc

**Frame works:**Flask, Django, Falcon, Bottle

**Tools:** Jupyter, Git, Jira, Docker

**Operating System:**Linux, Windows, MacOS

**Education:**

**California State University, East Bay**  2016 – 2018

**Master of Science, Business Analytics**

Relevant courses-Data warehousing & Business intelligence, Big Data Tech & Applications, Global Supply Chain, Data Mining.

**SRM University, India** 2010 – 2014

**Bachelor of Technology, Information Technology**

Relevant courses- IoT, Data Structures, Cloud Computing, Object-Oriented Analysis and Design.

**PROFESSIONAL EXPERIENCE**

**TCS/FEDEX September 2024- Present  
Sr. MLOps Engineer  
Responsibilities:**

* Streamlined the NLP workflow and automating lookup embedding file generation in both staging and production environments, ensuring the model consistently generates accurate and up-to-date codes using NLP. 
* Optimized workflow efficiency by transitioning from CPU-based to GPU clusters, reducing processing time by 25% and overall costs by 45%, showcasing the scalability and efficiency of MLOps practices in large-scale workflows. 
* Addressed resource inefficiencies by resolving issues such as running code on CPU in GPU clusters, switching to job-specific compute clusters, and fixing file naming and storage inconsistencies, leading to enhanced cost and time management in workflows. 
* Reduced processing costs by 25-45% and improved workflow sustainability for daily deployments by implementing best practices in resource allocation and cluster optimization. 
* Developed and implemented a robust framework for continuous model retraining in the NLP, enabling regular model updates to adapt to evolving shipment data trends.
* Integrated MLflow into the model retraining pipeline for automated tracking of key metrics such as accuracy and loss, ensuring consistent performance monitoring across all model versions.
* Cut the retraining processing time from 70 hours to 7 hours by leveraging GPU clusters, significantly improving the efficiency of regular retraining cycles while maintaining cost-effectiveness.
* Automated metric tracking with MLflow, eliminating manual errors and ensuring reliable performance evaluation, contributing to more accurate model promotions to production.
* Centralized experiment tracking and comparison of model performance using MLflow, facilitating faster iterations and enabling data-driven decision-making for selecting the best model versions.
* Improved reproducibility and streamlined workflows, reducing errors and ensuring that model performance met operational requirements.
* Optimized workflows for future growth by exploring parallelization, dynamic embedding generation, and model registry integration, preparing the system to handle increasing data complexity and dynamic updates.
* Implemented governance frameworks to ensure models in production were properly monitored and tracked, guaranteeing continued reliability and performance even as the system scaled.

**Environment:** Azure, GPU Clusters, MLflow, Kubernetes, Databricks, Python, Azure DevOps, Spark, Git, Azure Blob storage, CI/CD pipelines, NLP Machine Learning, Azure Cloud

**Ebay | San Jose, CA Dec 2021 – August 2024**

**Sr. Data Scientist / MLops Engineer**

**Responsibilities:**

* Constructively been part of a talented research team of data scientists in the field of **Computer Vision** to innovate, analyse application requirements into data models, to support standardization & effective adoption of bleeding-edge scientific norms and practices with a vision to enable integration and collaboration of **AI/ML** into everyday workflow.
* Implemented deep learning techniques Fully Convolutional Networks (**FCNs**), Convolutional Neural Networks (**CNNs**), and Deep Neural Networks (**DNNs**) for various computer vision tasks such as image classification, object detection, and semantic segmentation.
* Designed and implemented **CI/CD** pipelines for machine learning models using **Jenkins**, **GitLab** CI, and Azure **DevOps**, ensuring rapid and reliable model deployment.
* **Containerized** machine learning applications using **Docker** and deployed them on **Kubernetes** clusters, ensuring high availability and scalability.
* Implemented **monitoring** and **logging** solutions using **Prometheus**, **Grafana**, **ELK** **Stack**, and **Azure** **Monitor** to track model performance and infrastructure health.
* Utilized **MLflow** and **DVC** for model versioning, tracking, and reproducibility, ensuring consistent and reliable model deployments.
* Developed automated workflows for **model** **retraining** and **deployment** based on data drift and performance metrics using Apache **Airflow** and **Kubeflow** Pipelines.
* Integrated **LLM**-powered chatbots into eBay's messaging system to provide real-time responses to customer queries, improving response times and user satisfaction.
* Designed and implemented robust machine learning pipelines using **Kubernetes(K8)**/**AKS** with **Argo** Workflow orchestration, ensuring scalable and efficient end-to-end ML processes.
* Managed data pipelines for ML workflows on **Kubernetes(K8)/AKS** using Argo Workflows, ensuring efficient **data** **movement** and **transformation** between pipeline stages.
* Built and maintained end-to-end machine learning pipelines, from data ingestion to **model** **deployment**, using tools like Apache **Airflow**, **MLflow**, and **Kubeflow**.
* Led the development of **computer vision** algorithms to enable augmented reality (AR) functionalities on eBay's mobile app, allowing users to visualize products in their living spaces.
* Implemented object detection and tracking algorithms to accurately place virtual 3D models of products within real-world environments, enhancing the shopping experience on eBay.
* Developed **Spark** code using **Scala** and Spark-SQL for faster processing and testing, integrating **MLOps** practices for efficient development workflows.
* Performed data cleaning and feature selection using **MLlib** package in PySpark, working with deep learning frameworks such as **Caffe** with considerations for **MLOps**.
* Integrated **CI/CD** pipelines with **Argo** Workflows and **AKS** to automate the deployment of updated machine learning models, ensuring continuous delivery and integration.
* Created an algorithm that can predict the type of the object in a typical house using Deep Learning. Used OpenCV for the image analysis and **keras** and **Tensorflow** for implementing artificial neural networks (ANN).
* Designed and implemented end-to-end deployment automation for applications in Azure Kubernetes Service (**AKS**) using GitLab **CI/CD** and **Jenkins**, ensuring smooth and efficient CI/CD pipelines.
* Developed **Docker** images for application components, optimized **Dockerfiles** for efficient builds, and pushed images to container registries for use in AKS deployments.
* Utilized Kubernetes (**Kubectl**) for orchestrating the deployment, scaling, and management of containerized applications within **AKS**, ensuring high availability and reliability.
* Configured **GitLab** CI/CD pipelines to automate the building, testing, and deployment of applications to **AKS**, improving efficiency and reducing manual intervention.
* Integrated model deployment pipelines with **Databricks** Jobs and CI/CD processes, ensuring automated and reliable model deployment and updates.
* Engineered highly scalable backend **REST** **APIs** to efficiently collect and aggregate data from a Data Lake, ensuring seamless data flow for analytics and reporting.
* Led a project to increase the **Clickthrough rate** (CTR) of display ads on eBay using Logistic regression. Wrote complex **SQL** queries to get ML features for the project.

**Environment:**  Computer Vision, GCP Dataflow, Google BigQuery, Apache Nifi, Apache Beam, RDF Graphs, Python, Scala, Spark-SQL, JSON, XML, Matplotlib, ggplot2, MySQL databases, SalesForce, PySpark, Caffe, Azure Cog Search, AWS Cloud Search, OpenCV, keras, Tensorflow, NLTK, TextBlob, Spacy, Gensim, Hadoop, Hive, Linux, Azure tech stack, GCP Vertex AI, PyTorch, Azure, OpenAI, Hugging Face's Transformers, PowerBI

**Kaiser Permanente | Oakland, CA Jan 2020 – Jun 2021**

**Sr. Data Scientist**

**Responsibilities:**

* Applied Supervised Machine Learning Algorithms for the **predictive modelling** to tackle various types of problems: Successful Transition from Skilled nursing facility, identify predictors for medicare advantage members, lower the cost for mitigating homelessness, issues management.
* Built a datawarehouse by utilizing **ETL** processes with tools such as **Apache Nifi** and **Apache Hadoop** to gather all the business data related to doctors, patients, prescriptions, orders, and calls from different sources.
* Deployed **OpenAI's** advanced machine learning models to analyze patient data and improve healthcare outcomes at Kaiser, contributing to personalized treatment plans and predictive analytics.
* Deployed **LLMs** in patient interaction systems to enhance virtual assistants and chatbots, providing patients with accurate and natural language responses, improving communication, and offering personalized healthcare information.
* Developed doctor report cards for real-time insights into their performance over the years. Using **Apache** **Kafka** for data ingestion and **Tableau** integrated with **Hadoop**/**Spark** for creating the reports.
* Developed predictive models like disease risk, readmission risk using **advanced machine learning** algorithms, ensemble models, and deep learning architectures, integrating **MLOps** practices for model deployment and monitoring.
* Used **Pandas, NumPy, Scikit-learn** in **Python** for developing various like emergency department wait time, chronic disease progression machine learning models and utilized algorithms such as Linear regression, Logistic regression, Gradient Boosting, SVM and KNN, incorporating MLOps for efficient model development and deployment.
* Developed pipeline using **Hive** (**HQL**) to retrieve the data from **Hadoop** cluster, **SQL** queries to retrieve data from **MySQL** database and used **ETL** for data transformation.
* Derived data from relational databases to perform complex data manipulations and conducted extensive data checks to ensure data quality. Performed Data wrangling to clean, transform and reshape the data utilizing **NumPy** and **Pandas** library.
* Implemented model versioning and **A/B testing** strategies on **Databricks** for evaluating model performance and conducting experiments to improve model accuracy and effectiveness.
* Utilized **A/B testing** to refine the appointment scheduling interface within Kaiser's online platforms, improving usability and ensuring a seamless scheduling experience for patients.
* Utilized IOT sensors for collecting health information of cold storage and build streaming data pipeline into **GCP**’s **BigQuery** with the help of **Apache** **Airflow**.
* Employed **Docker** in the deployment of deep learning architectures, providing a consistent runtime environment for image analysis tasks using **OpenCV**, **Keras**, and **TensorFlow**.
* Productionized machine learning pipelines to gather data from **BigQuery** using **Apache** **Airflow** and build forecasting models to forecast and predict temperature and humidity spikes inside the cold storage.
* Built monitoring dash boards by employing visualization tools such as **Tableau** or **PowerBI**, visualizing both the current state and predictive health of cold storage warehouses.
* Leveraged **Kubeflow** pipelines to automate end-to-end machine learning workflows for Kaiser’s applications, enhancing reproducibility and scalability in analyzing medical data and deriving insights, incorporating MLOps for streamlined pipelines.
* Applied **NLP** techniques for sentiment analysis on customer feedback and reviews.
* Led the development and deployment of machine learning models on **GCP** **Vertex** **AI** tailored for Kaiser’s applications, including predictive analytics for patient outcomes and disease progression.
* Designed end-to-end machine learning pipelines on **GCP** **Vertex** **AI** with a focus on security and compliance, ensuring that data handling adheres to regulatory standards like HIPAA.
* Utilized **Vertex AI**'s **AutoML** capabilities to fine-tune models for medical image analysis, ensuring high accuracy in tasks such as radiology image interpretation.
* Integrated **Docker** and **Kubeflow** into Kaiser’s data science workflows, allowing for efficient collaboration and sharing of machine learning models and experiments within the research team, with a focus on improving analytics and MLOps practices.
* Led the development of Deep Learning models, utilizing **PyTorch** and **Tensorflow**, to address intricate challenges and enhance predictive capabilities.
* Leveraged **Python**, **PyTorch**, and **Tensorflow** to design and implement cutting-edge models, enhancing the organization's capabilities in applied research and data-driven decision-making.
* Applied **OpenAI**'s natural language processing capabilities to analysed and understand unstructured clinical notes.

**Environment:** AWS, GCP (Google Cloud Platform), GCP Dataflow, Google BigQuery, GCP Vertex AI, Vertex AI's AutoML, Apache Nifi, Apache Hadoop, Matplotlib, Seaborn, Python, Tableau, Apache Kafka, Hadoop/Spark, Pandas, NumPy, Scikit-learn, Hive (HQL), MySQL, IOT sensors, Apache Airflow, Tableau, PowerBI, NLP (Natural Language Processing), OpenAI's natural language processing capabilities, Hugging Face's Transformers library, PyTorch, Tensorflow, Azure tech stack, Azure.

**Capital One | McLean, VA Sep 2018 – Dec 2019**

**Machine Learning Consultant**

**Responsibilities:**

* Built an ML model to automate the process of finding the root cause over failed events on store self-checkout machines (POS systems). Integrated the ML model by utilizing **Flask API**.
* Decreased the enterprise service now tickets by 15% in building a service named Back up as a service by utilizing **AWS Backup** which gives the ability for a customer to initiate backups, restores on servers.
* Led customer data migration project of ETL codes from **SAS** to **Python** using Azure Databricks.
* Successfully transitioned legacy **SAS** scripts to **Python**, enhancing scalability, flexibility, and maintainability of data processing workflows at Capital One.
* Utilized **Python** programming language and cloud-native technologies to refactor **SAS** scripts, optimizing them for deployment on cloud platforms such as **AWS** and **Azure**.
* Constructed a machine learning model for Capacity Planning by collecting historical CPU and Disk usage data from on-premises infrastructure, preprocessing the data, engineering features, and selecting suitable algorithms, such as **LSTM** networks, to forecast resource utilization.
* Implemented automated data ingestion pipelines using cloud based ETL tools like **AWS** **Glue** or **Azure** **Data** **Factory** to streamline the process of extracting, transforming, and loading (ETL) data into the dashboard.
* Using **Apache** **Airflow**, built data pipelines to gather data from store-checkout devices into **BigQuery**.
* Used **classification techniques** including Random Forest and Logistic Regression to quantify the likelihood of each user referring.
* Utilizing **Cloud Composer**, **BigQuery** and **GCP** **storage buckets** on Google cloud platform productionized machine learning pipelines to performs data extraction, data cleaning, model training and updating the model on performance basis.
* Designed and developed an automation process, that helps enterprise to maintain common configurations and detect configuration drifts across the enterprise virtual infrastructure using **Docker**.
* Applied **mean-variance optimization algorithms**, such as **Markowitz portfolio theory**, using optimization libraries like **scipy.optimize** in Python, to construct efficient investment portfolios balancing risk and return.
* Implemented anomaly detection algorithms, such as isolation forests and autoencoders, with Python libraries like **scikit-learn** and **TensorFlow/Keras** to detect and prevent fraudulent activities in financial transactions.
* Developed credit risk models using **gradient boosting machine algorithms**, such as XGBoost and LightGBM, in **Python**, to assess creditworthiness and predict default probabilities for loan applicants.
* Present data using data visualization techniques by using tools like **Tableau** and libraries like matplotlib, ggplot2, seaborn by creating graphs, charts, or other visualizations to convey the Results of data analysis.
* Optimized model serving infrastructure on **Databricks** for low-latency inference, utilizing features such as model caching, distributed serving, and parallel processing.
* Utilized python and **Kafka** to build data pipelines for pulling data from multiple sources (vCenters, data bases, store devices) into Google’s **BigQuery**.
* Perform Quality Analysis testing and validation internally using **Django** and reformulate models to ensure accurate prediction of outcomes of interest and end to end **API** testing with dummy data and actual data before launch of the actual product with the Engineering team.
* Utilized machine learning algorithms such as logistic regression, multivariate regression, K-means, & Recommendation algorithms to extract the hidden information from the data.
* Used **Pandas**, **NumPy**, **Scikit-learn** in Python for developing various machine learning models and utilized algorithms such as Linear regression, Logistic regression, Gradient Boosting, SVM and KNN.
* For serving data, built **REST APIs** on the data lake (**BigQuery**, cloud **SQL**).
* Provided **Agile** coaching and training to teams, ensuring a common understanding of Agile principles, practices, and ceremonies for efficient project delivery.

**Environment:** Flask API, AWS Backup, LSTM networks, Apache Airflow, BigQuery, Random Forest, Logistic Regression, Cloud Composer, GCP Storage Buckets, Google Cloud Platform, Python Django framework, Cloud Bolt, Docker, Scikit-learn package, MS PowerBI, ggplot, Seaborn, Matplotlib, Kafka, REST APIs, Dataflow, Google App Engine, Linux servers.

**Mastercard | Purchase, NY Jun 2017 – Aug 2018**

**Data Scientist**

**Responsibilities:**

* **Identified business problems** or management objectives that can be addressed through data analysis and propose creative solutions and strategies to existing business challenges.
* Analyse, manipulate, and process massive amounts of data using statistical software to discover trends, patterns, and insights via Jupyter, Sci-kit learn and **Tableau**.
* Automate the entire collection process pipeline by identifying valuable data sources by using **ETL** tools like **Apache** **Nifi**, **Apache** **Beam**.
* Apply feature selection algorithms to models such as ANOVA (analysis of variance), decision trees using **PySpark’s** **Mllib** package and hyper tune the parameters based on interest and to predict the outcomes.
* Developed and maintained **Tableau** dashboards used by Transaction Monitoring, and FIU departments for reporting essential Anti-Money laundering (**AML**) transactional metrics, improving workstream by 17%.
* Revamped **PostgreSQL**/**SQL** to **Hive** queries, increasing processing speed by 40% and reducing deployment time by three weeks.
* Examined the behaviour of millions of Asset Transfer Party, Journal, and Trade Execution transactions from the L1 layer to the L4 layer in the Enterprise Analytics Platform to understand the attributes that impact Feature Calculation.
* Reviewed 120 AML Data Quality rules consistent with the Business Requirement Documents (BRD) and Functional requirement documents (FRD) and made updates based on the upstream rule change requirements.
* Experienced in handling large datasets using Partitions, spark in-memory capabilities, Broadcasts in spark, effective and efficient Joins, Transformations, and others during the ingestion process.
* Automated **data governance** for asset management data using **Python scripts**, reducing manual efforts by 60%.
* Optimized **SQL** queries on AML transaction monitoring rules by automating data cleansing, extraction, and analysis processes, resulting in improved performance and data accuracy by 50%.
* Incorporated data assessment by reviewing and monitoring **Anti-Money laundering** (AML) transactions consistent with global standards and procedures.
* Identified close to **100k fallout**s in the Data Quality report and executed **Hive** and **SQL** queries on the **Hadoop** platform to understand the root cause of the defect and underlying transaction and reference data.
* Assessed the 32 **Critical Data Elements** in the AML transactions responsible for NextGen and Feature Impact.
* Worked on optimizing **ELT** workloads against the **Hadoop** file system by implementing **Hive** **SQL** for transformation and performance tuning methodology in optimizing SQL, ETL mappings, and HIVE- HIVE-managed / ORC tables.
* Utilized version control systems like **Git** and collaboration platforms like **Jira** to facilitate seamless collaboration with cross-functional teams and track project progress efficiently.

**Environment:** Hadoop, Jupyter, Sci-kit learn, PySpark’s Mllib, Tableau, HDFS, Apache Nifi, Apache Beam, Shell Scripting, Python, Tableau, Hive SQL, SQL, Microsoft Excel, Business Analysis, BRD, FRD, HP ALM, Defect tracking, Feature Engineering, Git, Jira.

**Accenture, HYD Sep 2014 – Oct 2016**

**Data Engineer**

**Responsibilities:**

* Conducted qualitative and quantitative research to gather data from **data mart**.
* Responsible for data identification, collection, exploration & cleaning for **modelling**, participated in model development.
* Visualized, interpreted, report findings, and developed strategic uses of data.
* Understood transactional data and developed analytics insights using **Statistical models** using Machine learning.
* Involved in **gathering requirements** while uncovering and defining multiple dimensions. Extracted data from one or more source files and Databases.
* Collected database of sales of items in all aspects. Cleaned, filtered, and transformed data to specified format.
* Designed various intelligent reports using various reporting tools.
* Cleaned data using **R**, then visualized the data, and derived **statistical modelling** plots.
* Performed data visualization via **ggplot2** in **R** and **matplotlib** in **Python**.
* Worked in **Amazon Web Services** cloud computing environment.
* Responsible for **providing reports**, analysis, and insightful recommendations to business leaders on key performance metrics pertaining to sales & marketing.
* Gathered all the data that is required from multiple data sources and creating datasets that will be used in analysis.
* Used R to identify product performance via Classification, tree map and regression models along with visualizing data for interactive understanding and decision making.
* Created Intelligent dashboards and visualization on regular basis using **ggplot2** and **Tableau** **Tabpy**. Reserve.
* Accomplished multiple tasks from collecting data to organizing data and interpreting statistical information.
* Created dynamic linear models to perform trend analysis on customer transactional data in R.
* Conducted exploratory and descriptive data analysis of large data sets.
* Expertise in **Business Intelligence** and data visualization using **R** and **Tableau**.
* Performed Exploratory Data Analysis and Data Visualizations using **R**, and **Tableau**.
* Applied concepts of probability, distribution, and statistical inference on given dataset to unearth interesting findings through use of comparison, T-test, F-test, R-squared, P-value.

**Environment**: Data mart, Statistical models, Machine learning, R, ggplot2, Matplotlib, Amazon Web Services (AWS), Reporting tools, Classification models, Tree map models, Regression models, Tableau Tabpy, Business Intelligence, Exploratory Data Analysis (EDA)