**FULL NAME**

**(123) 456-7890 | email@mail.com**

**PERSONAL SUMMARY:**

AI/ML Engineer with over seven years of hands-on experience and comprehensive industry knowledge in the areas of – Machine Learning, Deep Learning, Data Mining, Statistical Modelling, Data Modelling, Data Management, Data warehouse, Data Visualization, solving Real-world practical problems. Experience in configuring reports from different data sources using data blending. Proficient in creating end to end Generative AI web application using Python, pyspark, vector database and Large Language Models (LLM).

**PROFESSIONAL SUMMARY:**

* Spearheading industry-specific innovations within Data Modernization and Generative AI, incorporating technologies like Open AI, Google BARD, and Azure.
* Hands-on Improvement helping clients in making and adjusting worksheets and information representation dashboards in Power BI.
* Reducing execution time in my projects significantly by performing extensive independent research for problem solving
* Proficient in creating end to end Generative AI web application using Python and Stream lit and Large Language Models (LLM) like Llama, GPT and Bert.
* Integrating and automating runs / workflows across multiple platforms seamlessly in both AWS, GCP. Have integrated multiple platforms like Snowflake, Databricks, AWS, Docker, ThoughtSpot etc. for the Data Science application to work seamlessly.
* Designing an AI chatbot using TensorFlow- Keras and deploying it in the cloud successfully.
* Working and scaling the big-data distributed data processing and discovering the hidden patterns.
* Experienced in working with popular NLP (NLTK, genism), LLM (Llama, Bert, GPT) and other frameworks like PySpark for developing and deploying natural language processing models.
* Proficient in developing ETL jobs and stored procedures for data extraction/Transformation/Loading from different data sources to target data warehouse.

**TECHNICAL SKILLS:**

|  |  |
| --- | --- |
| Languages | Python, PySpark, SQL |
| Databases & Frameworks | SQL, NoSQL, Hadoop, Spark, Huggingface, ZenML, MLflow |
| Cloud Technologies & Tools | GCP, AWS, Databricks, Snowflake |
| Processing Frameworks | Kafka, Spark, Databricks, Terraform |
| Visualization | Tableau, Power BI, Dash |
| Data Science Packages | Spacy, TensorFlow, Pytorch, NLTK, Keras, Pandas, NumPy, SciPy, Theano, Scikit- learn, Seaborn, Matplotlib, Plotly, dplyr, ggplot 2, Bokeh, Hugging Face, Open AI & Lang Chain |
| Machine Learning/DL Techniques | Regression, Classification & Clustering models, Random Forest, SVM, K-Nearest Neighbors, Bayesian, Neural networks (CNN, RNN, Feedforward), NLP (N-Grams, Markov model), Forward & Backward Propagation, Gradient Descent, Feature Scaling, Regularization, Normalization, ARIMA, LSTM, PCA, t-SNE. |
| Large Language Models/Techniques | Text embeddings, Attention mechanisms, Transformer model architecture, fine-tuning, RAG/Connectors, Semantic search, Endpoints, Evaluation Metrics, Text classifications, Text generation, Text representation, prompt engineering, Content moderation, Entity extraction, Topic Modelling, Sentiment analysis, GAN’s and VAE |
| Deployment Techniques | User interface: Streamlit, Data button.  Endpoints: FastAPI, Flask, RestAPI, Amazon SageMaker, API Gateway |
| Software Version Control & Documentation | Git, GitHub, GitLab, JIRA, Confluence, CI/CD |

**WORK EXPERIENCE:**

**AI/ML Engineer**  **Nov 2023 – Present**

NerdWallet, Remote.

* Designed advanced Generative AI (GEN AI) powered applications using Large Language Models (LLMs), NLP, Google Dialogflow, and Retrieval-Augmented Generation (RAG), improving customer interaction experiences by providing contextually relevant and accurate responses.
* Engineered end-to-end machine learning pipelines, integrating GPT (LLM) for natural language processing (NLP), natural language understanding (NLU), relevancy, and ranking in document retrieval, enhancing automated support and content generation workflows.
* Led the PoC of a scalable GPT (LLM)-based chatbot using Decoder Transformer architecture, Google Dialogflow, and LangChain, facilitating automated, human-like responses for customer support.
* Built and maintained real-time data pipelines using Apache Kafka and Google Cloud Pub/Sub, enabling low-latency streaming for machine learning models and improving response times for customer queries.
* Implemented document retrieval systems with NLP-based query expansion and semantic search, improving the relevancy and ranking of search results for customer support applications powered by Google Dialogflow for enhanced user interaction.
* Applied NLP techniques to analyze customer feedback and reviews, extracting key insights using sentiment analysis and topic modeling, helping product teams prioritize feature enhancements.
* Implemented Google AI Platform for training and deploying machine learning models, ensuring efficient handling of large datasets and automated scaling for prediction services, reducing model deployment time and enhancing relevancy in predictions.
* Applied RAG techniques, combining traditional machine learning and information retrieval systems, improving the ranking, relevancy, and accuracy of customer query resolution through Google Dialogflow.
* Collaborated with data science teams to integrate GPT-based and NLP models for content summarization, relevancy ranking, and user query resolution, significantly improving the product knowledgebase.
* Developed CI/CD pipelines using GitLab and Google Cloud Build to streamline the deployment of machine learning models, ensuring rapid iteration, versioning, and relevancy improvements across production environments.
* Enhanced model performance with distributed systems using Apache Spark on Google Cloud Dataproc, accelerating data processing times and enabling real-time model inference with accurate ranking.
* Collaborated with DevOps and SRE teams to ensure high availability and monitoring of AI systems, optimizing infrastructure using Google Kubernetes Engine (GKE) for orchestration and deployment.
* Integrated Google Identity Platform for secure, scalable user authentication in AI-driven customer support platforms using Google Dialogflow, enhancing data privacy and secure access to sensitive data.

**Environment**: LLMs, RAG, Google Dialogflow, Hugging Face, LangChain, Python, Google AI Platform, Apache Kafka, Google Cloud Pub/Sub, Google Cloud Dataproc, Spark, GitLab, Google Kubernetes Engine, Docker, Google Identity Platform, ELK Stack, CI/CD

**ML Engineer**  **Jan 2020 – Nov 2023**

Honda Motor Company, Remote.

* Architected and implemented a scalable real-time data pipeline using AWS Kinesis, Lambda, and S3 to process and onboard telemetry data for real-time analytics, ensuring relevancy of streaming data analysis.
* Developed data onboarding and processing workflows leveraging Managed Streaming for Apache Kafka (MSK) and PySpark to stream, enrich, and transform vehicle telemetry data, integrating NLP algorithms for automated text extraction and ranking of insights based on data importance.
* Architected the cloud infrastructure using AWS CloudFormation (IAC) to automate the provisioning of resources like S3, EC2, RDS, Redshift, etc., optimizing for high relevancy in data flow and processing.
* Integrated Snowflake as the central data warehouse, optimizing storage and query performance through partitioned tables, SQL-based transformations, and NLP-driven data categorization, ensuring faster retrieval and ranking of critical data.
* Collaborated with data scientists to design and deploy machine learning models using Python, PySpark, and SageMaker, applying ranking algorithms to prioritize vehicle component failure predictions and trigger proactive maintenance based on high-relevancy alerts.
* Applied AWS Glue Data Catalog to organize and manage metadata, utilizing NLP to classify and rank data for efficient access and analysis, ensuring relevant data is prioritized in workflows.
* Developed and maintained large-scale data pipelines using Apache Spark, MapReduce, and PySpark, with MSK handling big data processing, ensuring that high-relevancy data is ranked, for faster ETL workflows.
* Implemented supervised learning techniques, including Support Vector Machines (SVM) and Decision Trees, on Amazon SageMaker, with NLP techniques applied to classify and rank incoming data.
* Leveraged Scikit-learn, TensorFlow, and PySpark for training machine learning models on Amazon SageMaker, including NLP models for real-time text extraction and ranking based on relevancy to vehicle performance and user behavior.
* Developed and deployed MLOps pipelines in production utilizing SageMaker, SageMaker-Data Wrangler, Elastic Kubernetes Service (EKS), and PySpark, incorporating NLP-driven ranking systems to enhance data quality and relevance for real-time analysis.
* Managed the end-to-end data pipeline lifecycle, from architecture design to deployment and optimization, ensuring high availability and reliability of PySpark-based pipelines, with ranking algorithms to prioritize data streams based on their relevancy for downstream processes.
* Developed custom ETL processes using AWS Glue, PySpark, and SQL to transform raw data into actionable insights, incorporating NLP and ranking models to improve downstream.

**Environment:** AWS Kinesis, AWS Lambda, AWS S3, Apache Kafka, Snowflake, Python, AWS Glue, AWS CloudWatch, Power BI, Machine Learning, Data Pipeline, Terraform (Infrastructure as Code - IAC)

**Data Scientist**  **Mar 2017 – Nov 2019**

Southwest Airlines, Dallas, TX

* Collaborated with cross-functional teams to define dynamic pricing objectives and scope, integrating data from booking systems, sales, customer demographics, and competitor pricing into Snowflake for centralized processing.
* Developed a data pipeline using Google Cloud Dataflow to perform ETL operations, extracting and transforming high-volume datasets from disparate sources, ensuring scalability and efficiency.
* Conducted in-depth exploratory data analysis (EDA) using Pandas and NumPy, identifying data quality issues, correlations, and trends; generated feature insights through interactive Tableau dashboards.
* Engineered advanced features such as booking lead time, flight route popularity, historical pricing trends, and many more, leveraging domain knowledge to improve model accuracy and relevance.
* Developed machine learning models using TensorFlow and Scikit-learn, algorithms like Random Forest, Gradient Boosting for regression analysis, and LSTM, Prophet, SARIMA for time-series forecasting.
* Utilized K-fold cross-validation to evaluate model generalization, optimizing hyperparameters and selecting models based on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared scores.
* Built scalable, real-time prediction pipelines with Google AI Platform, implementing batch and online inference for dynamic pricing updates based on live data.
* Containerized machine learning models using Docker for portability, deploying via Google Kubernetes Engine (GKE) to manage clusters and automate scaling across multiple environments.
* Set up continuous integration/continuous deployment (CI/CD) pipelines to streamline model updates and ensure fast iterations using Git and Jenkins, improving deployment efficiency and accuracy.
* Integrated monitoring and alert systems via Google Cloud Monitoring to track model drift, system performance, and data anomalies, ensuring proactive maintenance and optimization.
* Boosted revenue per flight through real-time price optimization and improved customer engagement, contributing to more competitive and demand-driven ticket pricing strategies.

**Environment**: Python, SQL, PostgreSQL, Snowflake, Google Cloud Dataflow, TensorFlow, Scikit-learn, Pandas, NumPy, Tableau, Docker, Google Kubernetes Engine, Google AI Platform, Apache Spark, Google Cloud Monitoring

**EDUCATION:**

* University of Texas at Dallas - Master of Science in Business Analytics;   
  Graduate Certificate: Specialization in Applied Machine Learning.