Contact

132 Andrea ct, Lewisville, Texas - 75067
9062314017 (Home)
shaikrafi7@gmail.com

www.linkedin.com/in/rafi-shaik-33aa00100 (LinkedIn)

Top Skills

Support Vector Machine (SVM)

Agile Project Management

Exploratory Data Analysis

Languages

Hindi

English

Telugu

Certifications

Getting and Cleaning Data

Spark and Python for Big Data with PySpark

The Data Scientist’s Toolbox

PYTHON PROGRAMMING: INTERMEDIATE

Publications

Comparative Genomics

Polymorphisms and evolutionary history of retrotransposon insertions in rice promoters

Differential regulation of genes by retrotransposons in rice promoters

Classification method for microarray probe selection using sequence, thermodynamics and secondary structure parameters

Bioinformatic analysis of epigenetic and microRNA mediated regulation of drought responsive genes in rice

Patents

A system method and computer program product for pedigree analysis

Rafi Shaik

Data Scientist
Lewisville, Texas, United States

Summary

Multi-faceted professional with 10+ years of industry experience as product owner, lead architect, and data scientist. Ph.D. in Bioinformatics with 300+ citations

Experience

Verizon

1 year 5 months

Machine Learning Engineer

January 2025 - Present (1 year)
Irving, Texas, United States

Setup monitoring for performance, XAI (Explainable Artificial Intelligence), drift, stability, operations and custom metrics for a number of ML and deep learning models in production. Build a generic code template to generate all the common metrics such as lift, R2, RMSE, accuracy, precision and recall. Created script to perform SHAP (SHapley Additive exPlanations) analysis. Extensively used Airflow, GCP dataproc, bigquery, Jenkins and Gitlab to set up pipelines to load data. Created multiple Grafana dashboards with timeseries, histogram and bar charts, and customized layouts. Performed POC to create AI Agents to onboard new models to RT-AIMLOPs platform using VEGAS (Verizon Enterprise GenAI & Agentic Services).

Senior Data Scientist

August 2024 - Present (1 year 5 months)
Irving, Texas, United States

Project - Personalization AI (PzAI)

- Built, trained and tested models using PzAI architecture (a novel LLM based recommender system ref paper published by our team) for a number of use cases such as dynamic offers, purchase propensity and channel recommendations
- Improved model performance by fine tuning models on large datasets using A100/T4 systems on Domino Data Lab platform
- Wrote complex BigQuery SQLs for daily data ingestion and feeds to model training

- Performed analytics on large datasets using PySpark to identify key features and extract valuable insights relevant to use cases
- Implemented dynamic offer recommendation routing via Pega Adaptive Decision Manager (ADM)
- Build postman testing suite using postman scripts to run 100s of test cases on multiple API endpoints
- Collaborated with cross-functional teams to RCA a number of defects and followed through to closure
- Implemented model performance and drift monitoring in prod and retraining strategies.

Project - RT-AIMLOPs

- Setup monitoring for performance, XAI (Explainable Artificial Intelligence), drift, stability, operations and custom metrics for a number of ML and deep learning models in production
- Build a generic code template to generate all the common metrics such as lift, R2, RMSE, accuracy, precision and recall
- Created script to perform SHAP (SHapley Additive exPlanations) analysis
- Extensively used Airflow, GCP dataproc, bigquery, Jenkins and Gitlab to set up pipelines to load data
- Created multiple Grafana dashboards with timeseries, histogram and bar charts, and customized layouts
- Performed POC to create AI Agents to onboard new models to RT-AIMLOPs platform using VEGAS (Verizon Enterprise GenAI & Agentic Services)

AT&T

Senior Data Scientist

October 2023 - December 2025 (2 years 3 months)

Plano, Texas, United States

Fraud Detection using Knowledge Graph

Created knowledge graphs for various fraud detection use cases using Rel language on Relational AI platform

Performed EDA using PySpark, SQL and Python on large datasets from Azure and Snowflake in databricks notebooks

Defined complex ontology models using Microsoft Visual Studio

Generated graph export reports for number hijacking, post activation gaming and data mining fraud use cases

CVS Health

Data Scientist

October 2021 - June 2022 (9 months)
Dallas, Texas, United States

SAGE IT, INC

Data Scientist
October 2013 - June 2022 (8 years 9 months)
Frisco, Texas

CVS

Senior Data Scientist
January 2022 - May 2022 (5 months)

Developed Generalized Machine Learning Pipeline (GMP) in PySpark that ingests hive data and automatically does data cleaning, feature selection, model selection and tuning. Implemented a1c (target variable) missing value imputation using Multivariate Imputation by chained equations (MICE) algorithm. Developed deep learning models via Keras using Rx, Dx data to predict a1c. Co-ordinated development of a ML model to diagnose Asthma/ COPD in collaboration with an external client (Novartis). Built complex hive queries to collect required data from multiple sources. Developed clinical study design to test the ML model at CVS health hub locations.

Verizon

2 years 7 months

Senior Data Scientist
June 2019 - December 2021 (2 years 7 months)
United States

Analyzed user sessions data via clustering to auto-detect browsing patterns (workflows). Collate page URLs by user sessions for each application using New Relic API. Preprocess URLs via regex and convert URLs into bag of words using TF-IDF vectorization. Perform clustering using DBSCAN, OPTICS, K-Means. Segment user base for each application and identify potential issues/opportunities based on volume of partial workflows. Extract New Relic RUM data using NRQL (New Relic Query Language) for over 100 applications. Extract Catchpoint data using custom python script. Wrote complex regex to preprocess URLs and perform correlation, and store data in Postgres tables. Build tableau dashboard with custom visualizations (Venn diagram/Diverging bar chart). Build a new relic application (Nerdpack) to present health of databases via custom overall score and by segment such as Infrastructure, Application, Performance, Stability and Security. Wrote python scripts to query multiple New Relic entities to get metrics values and calculate custom scores. Wrote complex NRQL queries to provide derived metrics and recommendations such as Top long running queries, Blocking parent/child queries.

Lead Data Scientist

June 2019 - October 2021 (2 years 5 months)

Irving, Texas, United States

• Product owner of multiple data science/analytics projects actively contributing end to end including business requirement analysis, technical stack identification, complete solution & workflow design, POC demos, development, production support, and project management

• Proven track record in deriving business intelligence from complex data and delivering outstanding business value

• Designed end to end solution for detection of major network outages across Verizon networks around the globe for project Verizon Intelligent Outage Detection

• Orchestrated multiple rounds of data analysis and deep learning model building for the creation of the model with highest possible accuracy for project Managed Services Automated Repair Controller

AT&T

Data Scientist

December 2013 - April 2019 (5 years 5 months)

Dallas, Texas, United States

• Developed models to identify potential customers for new/related products based on user portal interactions, usage of product features, ordering history, by applying various classification algorithms such as multiple regression (linear/logistic), XGBoost, Deep Neural Network (Keras/PyTorch)

• Worked on more than 30 sprints of agile/SAFe project management.

• Provided end to end solution and workflows for interaction with enterprise clients

• Generated business process models, sequence diagrams, use case diagrams using Enterprise Architect

• Designed large scale APIs and models following OAS3.0 (Open API Specification) in SwaggerHub

• Performed data mapping across multiple interfaces generating comprehensive spreadsheets to maintain data integrity and provide one stop reference

• Regularly generated detailed documentation (Design documents, PPTs, Epics and User stories in Rally) of design solutions, system requirements/restrictions for development teams enabling accurate implementation

Michigan Technological University
Graduate Research Assistant
September 2009 - June 2013 (3 years 10 months)
Houghton, Michigan, United States

• Extensively worked on large volumes of gene expression data from disparate sources (Microarrays, Next-gen sequencers)
• Regularly performed data wrangling (cleaning, normalization, transformation and mining) and exploratory data analysis (EDA) activities
• Analyzed and classified multiple biological conditions using machine learning techniques PCA, PLS-DA, R-SVM and RF
• Published three highly cited research articles in top journals based on above work (Ref1, Ref2, Ref3, Ref4)
• Developed a custom web application to query and display curated data using JAVA, PhP and SQL
• Performed modular analysis of gene expression data using the R package and identified expression patterns of drought and bacterial stress in rice

Philips
Scientist
December 2006 - December 2008 (2 years 1 month)

Philips Research Asia
Scientist
November 2006 - November 2008 (2 years 1 month)

Methylation prediction of CpG islands in different tissues for bio-marker discovery using Perl and shell scripting. Developed enhanced algorithms for probe and primer sequence designs, utilizing SVM (Support Vector Machines) for probe classification and SVM-RFE for feature extraction (Ref). Filed a patent on data mining system to analyze family genetic history and predict disease conditions (Ref).

Education

Michigan Technological University
Doctor of Philosophy (Ph.D.), Bioinformatics · (2009 - 2013)

Andhra University
Master of Science (M.Sc.), Human/Medical Genetics · (2003 - 2005)

Andhra University

Master of Science, Human/Medical Genetics

IBAB

Postgraduate Degree, Bioinformatics

Michigan Technological University

Doctor of Philosophy, Bioinformatics