SHRUTIKA SUNIL PARAB

Data Scientist

sdata3202@gmail.com

732-444-3158

PROFESSIONAL SUMMARY

* Over 8 years of experience in Data Extraction, Data Modelling, Data Wrangling, Statistical Modeling, Data Mining, Machine Learning, and Data Visualization.
* Experienced in Data Cleaning, Model Training and Building, Model Testing, and Model Deployment.
* Experienced in Machine Learning algorithms like Linear, Logistic Regression, Decision Trees, Random Forest, KNN, and Naïve Bayes for both structured, semi-structured, and unstructured data.
* Worked on Data Visualization tools like Tableau and Google Analytics. Performed Statistical analysis on both descriptive and Predictive analysis using machine learning algorithms.
* Used Python Libraries like NumPy, Pandas, SciPy, Scikit-Learn, Matplotlib, and Seaborn.
* Experienced in SQL programming and creation of relational database models. Experienced in creating cutting-edge data processing algorithms to meet project demands.
* Specialized in Generative AI, Prompt Engineering, NLP, and LLM fine-tuning (OpenAI, BERT, and LLaMA).
* Involved in writing the complex structured queries using views, triggers, and joins. Worked with packages like Matplotlib, Seaborn, and pandas in Python.
* Experienced in Linear Regression, Logistic Regression, Random Forest, Decision Trees, Naïve Bayes, and K-Means.
* Worked in Current Techniques and Approaches in Natural Language Processing. Better Understanding of Statistical Analysis and Modeling, Algorithms, and Multivariate Analysis, and familiar with model selection, testing, comparison, and validations.
* Engineered contextual prompts using LangChain and Prompt Engineering strategies to enhance relevance, accuracy, and domain alignment of LLM responses.
* Experienced in Machine learning with NLP Text classification and prediction using Python.
* Worked with the Amazon Web Services Environment for database storage. Identify problems and provide solutions to business problems using data processing, data visualization.
* Experienced in Data Visualization using tables, lists, and tools like Tableau. Experienced in Business Intelligence tools like SSIS, SSRS, and ETL.
* Proficient in design and development of various Dashboards, Reports utilizing Tableau Visualizations like bar graphs, scatter plots, pie-charts, Geographic's and other making use of actions, local and global filters, cascading filters, context filters, Quick filters, parameters according to the end user requirements.
* Demonstrated success in optimizing clinical workflows, accelerating drug discovery insights, and enhancing software personalization through GenAI.
* Experience and knowledge in provisioning virtual clusters under the AWS cloud, which includes services like EC2, S3.
* Experienced in migration from heterogeneous sources, including Oracle to MS SQL Server. Experience in writing SQL queries and working with various databases (MS Access, MySQL and Oracle DB, PostgreSQL).
* Worked on Jupiter notebook, PySpark through Cloud Platform in EC2 instance using Putty and estimated models using Cross Validation, Log loss function, ROC curves used AUC for feature selection.
* Experience in data analytics, predictive analysis like Classification, Regression, Recommender Systems. Experienced in developing Custom reports and different types of Tabular Reports, Matrix Reports, Ad hoc reports, and distributed reports in multiple formats using SQL Server Reporting Services (SSRS).
* Adept at translating domain-specific requirements into scalable AI/ML solutions using Python, PyTorch, SQL, and REST APIs.
* Proven expertise in building RAG pipelines, deploying ML models via MLOps on AWS and Azure, and leveraging LangChain, Hugging Face Transformers, and Vector DBs (FAISS, Pinecone).
* Used Elastic Search in a NoSQL database to store the data, retrieve, and manage the documents in JSON format.
* Exposure to AI and Deep learning platforms/methodologies like TensorFlow, RNN, LSTM, and PyTorch.
* Experience in designing visualizations using Tableau software and publishing and presenting dashboards.

EDUCATION

MS in Artificial Intelligence

BS in Computer Engineering

TECHNICAL SKILLS

|  |  |
| --- | --- |
| Programming Languages | Python, R, SQL, Scala |
| Python Libraries | Pandas, NumPy, SciPy, Matplotlib, Seaborn, Scikit-learn, NLTK, SpaCy, PyTorch, TensorFlow, Keras, OpenCV, XGBoost, LoRA, MLlib |
| Machine Learning | Linear & Logistic Regression, Decision Trees, Random Forest, KNN, Naïve Bayes, SVM, XGBoost, PCA, Clustering (K-Means), Ensemble Methods (Bagging, Boosting), RAG |
| Deep Learning | ANN, RNN, LSTM, CNN, PyTorch, TensorFlow, Keras, MXNet, Caffe2, Theano, CNTK |
| NLP | Text Classification, Speech-to-Text, NER (Named Entity Recognition), OCR, Sentiment Analysis, Resume Parsing, Chatbots (api.ai, Dialogflow) |
| Big Data Tools | Apache Spark, PySpark, Spark Streaming, Kafka, MLlib |
| Databases | MySQL, MS SQL Server, Oracle DB, RDBMS, NoSQL (ElasticSearch) |
| Generative AI | LLM Fine-tuning (OpenAI, BERT, LLaMA), LangChain, Hugging Face, Prompt Engineering, RAG, FAISS, Pinecone |
| Data Engineering | ETL, Data Wrangling, Data Cleaning, Data Modeling, Feature Engineering, SQL Queries, JSON/XML, OLAP/OLTP |
| Cloud Platforms | AWS (EC2, S3, Lambda, EMR), Kubernetes (k8s), Amazon RDS |
| DevOps & CI/CD | RESTful APIs, CI/CD Pipelines |
| Visualization | Tableau, Power BI, HTML5/CSS for chatbot UI interfaces, Matplotlib, Seaborn |

PROFESSIONAL EXPERIENCE

Rakuten, New York, NY Oct 2024 - Present

Data Scientist

Responsibilities:

* Provided the architectural leadership in shaping strategic, business technology projects, with an emphasis on application architecture.
* Participated in all phases of data mining, data collection, data cleaning, developing models, validation, and visualization, and performed Gap analysis.
* Used Pandas, NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn, and 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 chatbot development initiative to improve customer interaction with the application.
* Developed the chatbot using api.ai.
* Designed and fine-tuned large language models (LLMs) for code generation and documentation tasks using prompt engineering best practices.
* Automated CSV to chatbot-friendly JSON transformation by writing NLP scripts to minimize development time by 20%.
* Conducted studies, rapid plots, and used advanced data mining and statistical modelling techniques to build a solution that optimizes the quality and performance of data.
* Presented AI roadmaps and proof-of-concept demos to stakeholders, aligning technical solutions with business goals.
* Demonstrated experience in design and implementation of Statistical models, Predictive models, enterprise data models, metadata solutions, and data life cycle management in both RDBMS, Big Data environments.
* Analyzed large data sets, applied machine learning techniques, and developed predictive models, statistical models, and developing and enhanced statistical models by leveraging best-in-class modeling techniques.
* Worked on database design, relational integrity constraints, OLAP, OLTP, Cubes, and Normalization (3NF) and De-normalization of the database.
* Orchestrated A/B tests and performance monitoring for generative models, optimizing prompts for accuracy and relevance.
* Collaborated with cross-functional teams to integrate AI-powered assistants into applications.
* Implemented deep learning algorithms such as Artificial Neural Network (ANN) and Recurrent Neural Network (RNN), tuned hyperparameters, and improved models with Python packages TensorFlow.
* Worked on customer segmentation using an unsupervised learning technique - clustering.
* Building Deep Learning Models using Keras, TensorFlow, and PyTorch for product recommendation and deploying models on k8 (Kubernetes) clusters.
* Implemented CI/CD workflows for AI models using GitLab CI and Docker, ensuring reproducible deployments.
* Utilized Spark, Kafka, Spark Streaming, MLlib, Python, a broad variety of Machine Learning Methods including classifications, regressions, dimensionality reduction, etc.
* Designed and implemented system architecture for Amazon EC2-based cloud-hosted solution for the client.

Environment: Python (Pandas, NumPy, Scikit-learn, NLTK, Matplotlib, Seaborn, SciPy), Spark, AWS (EC2), Kafka, Scala, HBase, MLlib, api.ai, OLAP/OLTP, GenAI, Prompt Engineer, PyTorch, TensorFlow, RDBMS, MapReduce

Organon, Jersey City, NJ May 2024 – Sep 2024

Data Scientist

Responsibilities:

* Compiled, defined, and analyzed data requirements for projects. Prepared and presented data reports and offered analytical and statistical interpretations using Machine Learning Algorithms.
* Performed data pre-processing and cleaning to prepare the data sets for further statistical analysis.
* Developed Machine Learning algorithms to find the number of health insurance claims and provided insights for even smarter healthcare.
* Involved in all phases like data collection, data cleaning, developing models, validation, visualization, and performing gap analysis.
* Performed Data Analysis using Pandas, NumPy, Seaborn, SciPy, Matplotlib, Scikit-learn, and NLTK, and developed various Machine Learning Algorithms such as Linear Regression, Multivariate Regression, Naïve Bayes, K-means, KNN, and Random Forest.
* Developed NLP pipelines to extract insights from batch records and SOP documents using prompt-tuned LLMs.
* Built a GenAI-driven document search assistant to retrieve historical CAPA and deviation trends across thousands of manufacturing records.
* Experimented with predictive models, including Logistic Regression, Support Vector Machine (SVC), Random Forest, provided by Scikit-Learn and XGBoost.
* Conducted workshops on prompt engineering and AI ethics for cross-functional manufacturing teams.
* Worked on different data formats like JSON and XML and performed different machine learning algorithms.
* Used Ensemble methods like Bagging, Boosting, Gradient Boosting Machines, and XGBoost Techniques to improve the accuracy for the weak learners.
* Used Bayesian Algorithm for Gaussian Naïve Bayes and Naïve Bayes for the prediction of new data from the training data.
* Collaborated with GxP compliance officers to ensure GenAI models followed audit traceability and validation SOPs.
* Used Natural Language Processing for dictating the documentation and translating the speech to text in the healthcare industry.
* Worked on large datasets, acquired data and cleaned the data, and analyzed trends by making visualizations using Matplotlib. Created reports to show the insights, helped to make correct decisions that advance patient care, showed the customers graphs and reduced the price, and improved health.
* Used Spark platform for analysis using PySpark library and performed splitting of date into clusters on AWS.
* Designed and developed the new interface elements and objects as required and Set Analysis to provide functionality using Tableau.
* Visualized the data with Graphs and Reports using Matplotlib, Seaborn, and Pandas packages in Python on datasets for analytical models to know the missing values, correlation between the features, and outliers.

Environment: Python (Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib, NLTK, SciPy), PySpark, XGBoost, AWS Lambda, Tableau, JSON/XML, GenAI, Prompt Engineer, Machine Learning, NLP

Reliance Jio Platforms Limited, Mumbai, India Oct 2018 – Jan 2024

Data Scientist

Responsibilities:

* Designed and deployed a VIA system to assess 1,000+ sales interviews by extracting multimodal features (sentiment, emotion, eye-tracking, transcripts), reducing manual screening time by 70%.
* Developed a document processing automation pipeline using OCR and NER models to extract, validate, and integrate candidate data with recruitment systems via REST APIs.
* Fine-tuned transformer-based resume parsing models using LoRA for lightweight inference, achieving 94% extraction accuracy across key candidate entities.
* Integrated RAG-based recommender systems for job-candidate matching, leveraging SpaCy, Doccano, and deep learning NER models.
* Built scalable ETL pipelines for model training and evaluation using Python and SQL, with clear technical documentation and model deployment strategies.
* Conducted predictive modeling and statistical analysis on large datasets to support user behavior analytics, improving engagement strategies by 15%.
* Implemented prompt engineering strategies to extract structured data from unstructured EMR text.
* Created interactive Matplotlib dashboards to visualize KPIs and model performance metrics for business stakeholders.
* Engineered a face finder system using InsightFace and Elasticsearch to detect and retrieve individual faces from event photos, aiding attendance analytics.
* Applied OpenCV and RESTful APIs for facial recognition integration, ensuring robust backend and frontend synchronization with CI/CD pipelines.
* Integrated GenAI-based symptom checker and triage assistant with prompt tuning using anonymized EMR datasets.
* Led chatbot development using Dialogflow and NLP models to automate user engagement across platforms, improving user retention by 20%.
* Developing optimized and scalable Machine Learning Algorithms capable of performing predictive modeling using PyTorch, Spark, SparkML, Scikit-learn, and other packages.
* Applied GenAI to generate synthetic yet realistic patient records for data augmentation and model testing.
* Designed lightweight HTML5/CSS chatbot interfaces with backend ML integration, tracking real-time metrics through Power BI dashboards.
* Conducted multi-GPU model training to accelerate vision-based model performance for document analysis and facial recognition.
* Involved in Deep Learning frameworks like TensorFlow, Theano, CNTK, and Keras to help customers build DL models.
* Collaborated with cross-functional teams (Product, Ops, and Analytics) to align ML pipelines with business requirements and user experience goals.
* Optimized analytics workflows using Python, Pandas, and NumPy, ensuring reproducibility and high model performance across environments.

Environment: Python, SQL, OpenCV, NumPy, Pandas, Matplotlib, Power BI, SpaCy, Doccano, InsightFace, Elasticsearch, Dialogflow, HTML5/CSS, REST APIs, CI/CD, GenAI, Prompt Engineer, LoRA, RAG, Multi-GPU, PyTorch, Transformer Models

Zycus, Mumbai, India Mar 2017 – Sep 2018

Data Scientist

Responsibilities:

* Collaborated with the manager to gather all the information needed for data analysis and databases, and analyzed the raw data.
* Implemented and designed predictive models using Natural Processing Language Techniques and machine learning algorithms such as linear, logistic, and multivariate regression, random forests, k-means clustering, KNN, and PCA for data analysis.
* Involved in all aspects like data collection, data cleaning, developing models, and visualization.
* Maintained large data sets, combining datasets from various sources like Excel and SQL Queries. Writing SQL Scripts to select the data from the servers and modify the data as needed of Python Pandas and store it back to the different database servers.
* Created action filters, parameters, and calculated sets for dashboards and worksheets in Tableau. Published customized reports and dashboards, and report scheduling using Tableau Server.
* Performed data cleaning, exploratory analysis, and feature engineering using R. Performed data visualization with Tableau and generated the findings and enhanced customer satisfaction.
* Programmed in Python, which is used in packages like NumPy, Pandas, and SciPy. Developed the content involving data manipulation, visualization, Machine Learning, and SQL.
* Used different kinds of statistical models like Chi-Square Test, Hypothesis Testing, ANOVA, Correlation Testing, and Descriptive Testing.
* Implemented classification using supervised algorithms like Decision trees, KNN, Logistic Regression, and Naive Bayes.
* Understanding and analyzing the data using appropriate statistical models to generate insights.

Environment: Python, R, SQL, Tableau, Excel, Jupyter Notebook, Scikit-Learn, Pandas, NumPy, SciPy, NLP, ML Algorithms, Statistical Models