Pritsam Dabre

Data Scientist

(760)-282-6422 | pritsamdabre@gmail.com |github.com/Pritsam2| San Marcos, CA (Ready to Relocate)

SUMMARY

Data Scientist with 3+ years of experience in designing and deploying end-to-end machine learning solutions, including data preprocessing, model development, optimization, and deployment. Adept at handling both structured and unstructured data using Python, SQL, and cloud platforms like AWS and GCP. Experienced in real-time data stream processing with Kafka and scalable computing on Spark and EMR. Recently worked on projects involving Large Language Models (LLMs), Generative AI, and (RAG) using tools like FAISS, and Lang Chain. Passionate about solving business problems with data-driven insights and delivering impactful solutions in cross-functional environments.

WORK EXPERIENCE

S&P global, USA | Data Scientist. — March 2024 – Current

- Collaborated with cross-functional teams to design and deploy predictive models (e.g., logistic regression, XGBoost, neural networks, ARIMA) addressing business-specific challenges and improving decision-making accuracy.
- Wrote efficient Python and SQL scripts for large-scale data preprocessing, feature engineering, and statistical analysis, streamlining modeling workflows by 40%.
- Built and maintained ETL pipelines using PySpark, Apache Kafka, and Spark Streaming for processing structured and unstructured datasets, enhancing pipeline scalability and real-time analytics.
- Utilized cloud platforms such as AWS (EC2, EMR, S3, CloudWatch), Snowflake, and Google Cloud to store, process, and analyze high-volume datasets with cost-effective infrastructure.
- Developed dashboards in Tableau and Power BI for data storytelling and business insights, reducing manual reporting efforts and turnaround time by 40%.
- Integrated vector databases (FAISS) to enhance retrieval systems powering LLMs, improving relevance scores and user satisfaction by 35%.
- Applied A/B testing frameworks and statistical analysis to evaluate model and feature performance, increasing data-driven adoption by 18%.
- Deployed models using ML Ops tools including Git, Jenkins, and Docker, enabling CI/CD workflows that cut model deployment time in half.
- Designed automated alerting and monitoring systems using AWS CloudWatch to ensure pipeline reliability, improving uptime by 30% and reducing incident response time.
- Partnered with business stakeholders to gather requirements and translate them into actionable data science solutions aligned with product goals.
- Implemented machine learning pipelines using traditional and advanced models such as logistic regression, random forest, neural networks, NLP techniques, k-means clustering, ARIMA, and Prophet to support diverse use cases in forecasting, classification, and clustering.
- Authored technical documentation and presented model results to both technical and non-technical audiences, improving transparency and cross-team alignment.
- Designed automated monitoring and alerting systems using AWS CloudWatch, increasing data pipeline uptime by 30% and reducing incident response time.

Maruti TechLabs, India | Data Scientist — April 2020 - July 2022

- Specialized in predictive modeling, particularly NLP, to drive data-driven decision-making and optimize business outcomes.
- Led end-to-end data science projects, from data collection to modeling and visualization, delivering actionable insights.
- Extensively utilized Python libraries (Pandas, NumPy, Matplotlib, Seaborn, SciPy, Scikit-learn, NLTK) to streamline data processing and model development, increasing efficiency by 30%.
- Applied various machine learning algorithms and statistical models (decision trees, text analytics, NLP, regression models, SVM, clustering) to extract insights from large datasets, improving model accuracy by up to 15%.
- Implemented machine learning models (logistic regression, XGBoost, SVM) with Python's Scikit-learn, achieving up to 92% predictive accuracy on key business metrics.
- Integrated Power BI with Excel and other data sources, automating report generation and reducing manual reporting time by 40%.
- Leveraged AWS EC2 for scalable computing, optimizing data processing tasks and reducing model training time by 50%.
- Developed machine learning algorithms (linear regression, classification, Naive Bayes, Random Forest, K-means clustering, KNN, PCA) to perform complex data analysis, boosting business KPIs by 10%.
- Built and implemented predictive models using machine learning techniques such as linear regression, classification, multivariate regression, Naive Bayes, Random Forest, K-means clustering, KNN, PCA, and regularization for in-depth data analysis.
- Developed regression models, including Lasso, Ridge, SVR to accurately predict Customer Lifetime Value.
- Implemented backup and recovery strategies for SQL Server to ensure data availability and integrity during system failures.
- Improved marketing campaign ROI by 25% by developing machine learning models that predicted customer purchase behavior, leading to more targeted campaigns.

SKILLS

Programming Languages: Python, R, SQL, Scala, Julia, MATLAB, Java, Bash.
Machine Learning: Regression, Classification, KNN, Decision Trees, Random Forest, SVM, Naive Bayes, XGBoost, LightGBM, K-Means, Bayesian Methods, Sentiment Analysis, NLP, LLMs, CNNs, Neural Networks.
Deep Learning: TensorFlow, PyTorch, Keras, RNNs, LSTMs, GANs, Transformer Models (e.g., BERT, GPT), Hugging Face.
Data Visualization: Tableau, Power BI, Matplotlib, Seaborn.
Data Engineering & Analysis: EDA, Data Cleaning/Wrangling, Feature Engineering, Clustering, Forecasting, A/B Testing, Predictive Modeling, Time Series, Statistical Analysis, Experiment Design, Data Mining, Pattern Recognition, Data Integrity.
Databases: MySQL, PostgreSQL, MongoDB, Cassandra, Redshift, Snowflake, BigQuery.
Cloud & Tools: AWS (S3, EC2, Sage-Maker), GCP, Azure, Databricks, Apache Spark, Apache Kafka, Airflow, Docker, Kubernetes, Terraform, Jenkins, GitHub, Bitbucket.
Software Engineering & DevOps: OOP, Agile, CI/CD, DevOps, API Development, Model Deployment Version Control & Other Tools: Git, GitHub, Bitbucket, Excel (Advanced), SAP, JMP, Alteryx, Hadoop Operating Systems: Windows, Linux.
Soft Skills & Additional Competencies: Strong Communication Skills, Basic Leadership, Research-Oriented, Software Engineering Mindset, Problem-Solving.

EDUCATION

Master of Science in Computer Science (MS) | California State University - San Marcos, San Marcos, CA **CGPA: 3.8**

Relevant Coursework: Machine Learning • Artificial Intelligence • Artificial Neural Networks and Forecasting • Data Communication & Computer Networks • Advanced IoT • Introduction to Data Mining • Design Patterns and Object-Oriented Analysis

Bachelor of Engineering in Information Technology (BE) | Mumbai University- Mumbai, India **CGPA: 3.7**

Relevant Coursework: Design & Analysis of Algorithms • Database Management Systems • Data Structures using C++ • Object-Oriented Programming • Cloud Computing • Operating Systems

ACADAMIC PROJECTS

Content-Based Movie Recommender System (Python, scikit-learn, difflib, Streamlit, Heroku)

- Developed a content-based movie recommendation system using Python, scikit-learn, and difflib, analyzing movie metadata such as title, genre, and keywords to compute similarity scores.
- Built an interactive frontend using Streamlit, allowing users to input a movie and receive top-N similar recommendations with clean, responsive UI.
- Integrated cosine similarity and fuzzy matching techniques to improve the relevance of suggestions based on user queries.
- Deployed the application on Heroku, making it publicly accessible without local setup.

Email Spam Classifier – Machine Learning Web App (Python, Flask, Naïve Bayes)

- Developed a full-stack ML web app to classify emails as spam or ham, using Flask for backend and HTML/CSS for the frontend UI.
- Performed extensive feature engineering and data preprocessing, including tokenization, stopword removal, and vectorization with Count Vectorizer.
- Compared multiple models (Logistic Regression, SVM, Naïve Bayes) and selected Multinomial Naïve Bayes based on performance and efficiency; visualized key dataset insights using Matplotlib and Seaborn.
- Built a modular ML pipeline and deployed the app on PythonAnywhere, enabling real-time predictions through a live, user-friendly web interface.

End-to-End Image Classification Pipeline with Convolutional Neural Networks

- Built a modular ML pipeline and deployed the app on PythonAnywhere, enabling real-time predictions through a live, user-friendly web
- Developed a custom Convolutional Neural Network (CNN) to perform multi-class image classification, showcasing applied deep learning and computer vision skills.
- Built and trained the model using TensorFlow and Keras, with a focus on optimizing accuracy through experimentation with layers, activation functions, and learning rates.
- Preprocessed and augmented the dataset using resizing, normalization, and transformations like flips and rotations to enhance model robustness.
- Deployed the trained model with a command-line interface, enabling easy prediction on new images without requiring a graphical interface.

CERTIFICATIONS

Docker & Kubernetes – The Practical Guide (Udemy) ,**IOT-Certification**-National Business Idea Competition Winner – Pragati E Cell, **2022**
Introduction to Python Programming (Udemy) ,**The Complete SQL and MySQL Course** (Udemy),**The Spark Foundation** :Data Internship Certification(Spark Foundation), **Master Git and GitHub**: Go from Zero to Hero.