SaiReddy Thatiparthi

Data Scientist

Email: saiprakashreddythatiparthi@gmail.com

LinkedIn id:www.linkedin.com/in/saiprakash195

Phone: +14125405709

PROFESSIONAL SUMMARY:

- **4+years** of Experienced Data Scientist with a strong background in **Statistical analysis, Predictive Modelling, Data Cleaning, Data Visualization Data and Text Mining, Machine learning, Natural Language Processing (NLP), Deep learning Neural Networks**
- Skilled in using **Python** and tools like **TensorFlow**, **Scikit-learn** and **PyTorch** to build models and extract insights from large datasets
- Proficient in using **Python**, **Pandas**, **NumPy**, **Matplotlib** and **PowerBI** extract actionable insights from complex datasets, driving data-driven decision-making and business outcomes
- Skilled in using **SQL** to design and optimize complex queries, manage large databases, and extract valuable insights, ensuring efficient data retrieval and supporting data-driven decision-making processes.
- Aims to leverage expertise in data science to drive impactful business decisions in a dynamic environment
- Proficient in using **Databricks** notebooks for data exploration and model development, as well as Databricks Delta for efficient data versioning and storage
- Extensive experience in Python scripting for data manipulation, analysis, and modeling
- Data mining, and exploratory data analysis (**EDA**) to uncover hidden patterns and trends in structured, unstructured and Semi-structured data
- Proficient in deploying **Spark clusters** on AWS EMR for large-scale data processing
- Strong background in **statistical modeling** and **hypothesis testing**, using advanced techniques like **clustering**, **decision trees**, and **ensemble** methods to solve business challenges
- Strong skills in **Machine learning Algorithms**, **Deep learning Neural Networks**, and **Natural language processing** (**NLP**) models are deploying real-world applications
- strong Understanding of various neural Network types, including **ANN**, **CNN**, **RNN** and advanced techniques such as **Boltzmann Machine**, **Autoencoders** and **Generative Adversarial Networks** (**GAN**)
- Skilled in developing **AI**-powered solutions for tasks such as sentiment analysis, recommendation systems, and anomaly detection
- Experienced in working with AWS services such as **Amazon S3, EMR**, **Glue and AWS Lambda** functions for building data pipelines
- Proficient in leveraging **AWS SageMaker** to build, train, and deploy scalable machine learning models, enabling efficient and cost-effective AI solutions in production environments
- Proficient in exporting **SAS** results to diverse formats, including **XML, PDF**, and **Excel**, utilizing SAS/Export and SAS/ODS to create dynamic and visually compelling reports for effective data presentation and decision-making

EDUCATIONAL DETAILS:

Trine University, Angola, IN, USA **Dec 2023**

Master of science in Information Technology (Advanced Database, Data Mining, Data Visualization, Data Science and Big Data, Statistics & Quantitative Methods) GPA – 3.9/4.0

Universal College of Engineering and Technology, Guntur, AP **May 2020**

Bachelor of Technology in Mechanical Engineering

TECHNICAL SKILL SETS:

Programming Languages	**Python, Java, SQL**
Databases	**MySQL, MongoDB, SQL Server, DB2**
Hadoop Ecosystem	**Hadoop, HDFS, MapReduce, Hive, Impala, Pig, Sqoop, Oozie, Zena, Zeke Scheduling, Zookeeper, Flume, Kafka, Spark core, Spark streaming**
Web Frameworks	**Django,**
Web Technologies	**CSS,HTML5,Bootstrap,JQuery,ReactJS,JavaScript, NodeJS, Angular, Beautifulsoup**
Data Science & Machine Learning	**Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, PowerBI, Plotly**
Deep Learning Frameworks	**Tensorflow, Keras, PyTorch**
CI/CD Tools	**Jenkins, GIT**
Cloud Technologies	**Amazon Web Services(AWS)- S3, EC2, RDS, SageMaker**

AI/ML Models	**Regression, Classification, Time Series forecasting, Neural Network, Collabarative filtering, Deep learning, ANN, CNN, RNN**
Natural Language Processing	**SpaCy, NLTK, RASA, dialogflow**
Containers	**Docker and Kubernetes**
IDE	**Anaconda Navigator, Jupyter Notebook, Google Colab, VSCode**

CERTIFICATE:

Python Life **February 2019**

Full Stack Data Science

Edureka

Generative AI **January 2024**

Udemy

NLP - Natural Language Processing with Python

Certificate url: http://ude.my/UC-849a6ec6-7e12-4a06-acfb-154cc85df687

Udemy

Machine Learning Practical Workout| 8 Real-World Projects

Certificate url: http://ude.my/UC-5de09420-44d9-4452-a970-6659d44b9f7b

PROFESSIONAL EXPERIENCE:

American Airlines – Dallas, Texas **January 2024 - Present**

Data Scientist

Responsibilities:

- Performed **Data Collection, Data Cleaning, Data Visualization**, and Deep Feature Synthesis and extracted key statistical findings to develop business strategies.
- Analyzed loan application datasets to predict loan approval decisions using classification models **Logistic Regression**, **Decision Trees** and **Random forest**
- Filtered wanted and genuine data for extracting insight and for visualization of log data using Data Wrangler
- Implemented sampling, **PCA** and **LDA** for high dimensional data and drew visual statistical conclusions as well as **statistical inferences**.
- Performed **A/B testing** to involve in feature engineering and extracted useful feature without dependences on other features
- Analyzed and grouped products into different clusters based on **product description**, purchase and historic data using **k-means clustering technique**.
- Worked on analyzing different big data analytic tools including **Hive**, **Impala** and **Sqoop** in importing data from **RDBMS** to **HDFS**
- Involved in converting MapReduce programs into **Spark** transformations using **Spark** RDD's using **Scala** and Python.
- Trained a gradient boosted **Decision Tree** Regression and Classifier using **XGBoost** to predict actual customers lifetime value, aiding in strategic planning
- Developed and deployed to production multiple projects in the **CI/CD pipeline** for real-time data distribution, **storage**, and analytics Persistence to **S3**
- Wrote complex **SQL** and **PL/SQL** queries for stored procedures
- Used Cloudera Manager for installation and management of **Hadoop** Cluster
- Leveraged **AWS SageMaker** to build, train and deploy advanced **machine learning** and **deep learning** models, improving predictive accuracy and operational efficiency
- Integrated **Kafka-Spark** streaming for high efficiency throughput and reliability
- Utilized **Matplotlib** for data visualization to uncover new insights from extracted data, enabling clearer and more impactful representation of complex information
- Implemented the final machine learning models to integration and deployment using Containers like **Docker** and **Kubernetes**
- Created **data pipelines** for different events to load the data from **DynamoDB to AWS S3 bucket** and then into **HDFS location**

Environments: Hive, Sqoop, Storm, Kafka, HDFS, AWS, Data mapping, EC2, S3, Hadoop, YARN, MapReduce, RDBMS, Data Lake, Python, Scala, Databricks, Python Scripting, Dynamo DB, Pig, MongoDB, VSCode, NLTK, SpaCy

Infosys, India

Data Scientist **June2020-August2022**

Responsibilities:

- Utilized domain knowledge and application portfolio knowledge to play a key role in defining the future state of large, business technology programs**.**
- Collaborated with various teams to develop an analytics-based solution targeting roaming subscribers, enhancing customer engagement strategies
- Worked on data cleaning and ensured data quality, consistency, integrity using **Pandas, NumPy**
- Explored and analyzed the customer specific features by using **Matplotlib, Seaborn** in **Python** and **dashboards** in **Tableau**.
- Performed data imputation using **Scikit-learn** package in **Python**
- Participated in **features engineering** such as **feature generating**, **PCA**, **feature normalization a**nd **label encoding** with **Scikit-learn** pre-processing
- Demonstrated experience in design and implementation of Statistical models, Predictive models, enterprise data model, metadata solution and data life cycle management in both **RDBMS**, **Big Data** environments
- Designed and implemented system architecture for **Amazon EC2** based **cloud-hosted solution** for the client
- Enhanced an existing recommendation system using collaborative filtering algorithms in Python
- Worked on different data formats such as **JSON, XML** and performed machine learning algorithms in Python
- Reduced false positive rates by 15% through continuous model optimization and fine-tuning

Environment: Python 3.6, Spark, Hadoop, Kafka, PIG, HIVE, MatplotLib, Scikit-Learn, MySQL, SQL, Data Warehouse, Data Modeling, Middleware Integration, Gradient Boost, AWS SageMaker Random Forest, xgboost, OpenCV, Scikit-learn etc.

Hexaware, India **May 2019 -May2020**

Junior Data Scientist

Responsibilities:

- Analyzed market research survey data conducted by Cambridge research using **Pandas** help contributing to the development of a marketing and sales tool
- Performed exploratory data analysis(**EDA**) and statistical testing to understand data trends and provide actionable business insights
- Improved fraud prediction performance by using **random forest** and **gradient boosting** for feature selection with **Python Scikit-learn**.
- Led training sessions for tech teams in Canada, Australia, and the UK to ensure understanding of business logic and synchronization of reports across all three regions
- Created 30 individual reports and replicated them on **Matplotlib** dashboards, enhancing data visualization and accessibility for stakeholders
- Performed cluster analysis on grouping the customers based upon 56 variables using **K-mean cluster** analysis.
- Built models using **Statistical techniques** like **Bayesian HMM** and **Machine Learning** classification models like **SVM**, and Random Forest using with **Python** packages
- Developed a chatbot using natural language understanding **(NLU**) techniques with the Python **NLTK** library, improving customer interaction efficiency
- Used Natural Language Processing (**NLP**) for response modeling and fraud detection efforts for credit cards, reducing fraudulent activities and enhancing security measures
- Created complex and reusable Macros and used **SAS** and **Python** functions for Data Cleaning, Validation, Analysis and Report generation as the data variables
- Utilized Scikit-learn in Python to build predictive models for customer churn prediction, achieving a prediction accuracy of over 85% and enabling proactive customer retention strategies

Environment: Python, TensorFlow, Scikit-Learn, SQL, MySQL, AWS SageMaker, NumPy, Pandas, Matplotlib, Machine learning, Jupyter Notebook, Data Mining, Databricks, Visual analytics, SAS, NLTK, VsCode