**SAITEJA SIRIKONDA**

**Senior Data Scientist**

saitejasirikonda5@gmail.com|707-590-9664

**PROFESSIONAL SUMMARY:**

* Highly experienced Data Scientist with over 10+ years’ experience in Data Extraction, Data Modelling, Data Engineering, Statistical Modeling, Data Mining, Machine Learning and Data Visualization.
* Experienced in Software Engineering, and its intricacies – unit testing, CI/CD, pipelines, integration testing, code reviews, on-call shifts, customer interaction, Back-end API development, privacy and security concerns, being a scrum master and bug fixes.
* Domain knowledge and experience in Retail, Banking, Healthcare & Education industries.
* Expertise in transforming business resources and requirements into manageable data formats and analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across a massive volume of structured and unstructured data.
* Proficient in managing entire data science project life cycle and actively involved in all the phases of project life cycle including data acquisition, data cleaning, data engineering, features scaling, features engineering, statistical modeling, testing and validation and data visualization.
* Familiar with various cloud environments like AWS – EC2, S3, Lambda, Kinesis, Sage Maker, Quick Sight, DynamoDB etc. and also GCP - Big Query, Cloud Storage, Data prep, Cloud Composer (managed Apache Airflow), Data Studio.
* Proficient in Machine Learning algorithm and Predictive Modeling including Regression Models, Decision Tree, Random Forests, Sentiment Analysis, Naïve Bayes Classifier, SVM, Ensemble Models.
* Proficient in Statistical Methodologies including Hypothetical Testing, ANOVA, Time Series, Principal Component Analysis, Factor Analysis, Cluster Analysis, Discriminant Analysis.
* Knowledge on time series analysis using AR, MA, ARIMA, GARCH and ARCH model.
* Knowledge on Natural Language Processing (NLP) algorithm, Text Mining, Generative AI and LLMs.
* Worked in large scale database environment like Hadoop and MapReduce, with working mechanism of Hadoop clusters, nodes and Hadoop Distributed File System (HDFS).
* Collaborate with QA teams to perform integration testing.
* Understand and analyze the functional requirements of the software.
* Writing SQL queries within PL/SQL to retrieve, update, insert, and delete data in the database.
* Strong experience with Python (2.x,3.x) to develop analytic models and solutions.
* Proficient in Python 2.x/3.x with SciPy Stack packages including NumPy, Pandas, SciPy, Matplotlib and I Python.
* Develop data processing pipelines for large volumes of data using Hadoop and GCP - Set up a Hadoop cluster using Google Dataproc, design ETL processes to process large data sets & Load results into Big Query for analytics.
* Working experience in Hadoop ecosystem and Apache Spark framework such as HDFS, MapReduce, HiveQL, Spark SQL, PySpark.
* Very good experience and knowledge in provisioning virtual clusters under AWS cloud which includes services like EC2, S3, and EMR.
* Proficient in data visualization tools such as Tableau, Python Matplotlib, R Shiny to create visually powerful and actionable interactive reports and dashboards.
* Excellent Tableau Developer, expertise in building, publishing customized interactive reports and dashboards with customized parameters and user - filters using Tableau (9.x/10.x).
* Experienced in Agile methodology and SCRUM process.

**EDUCATION:**

* Bachelors in Computer Science, Indian Institute of Technology, Mandi-2014

**TECHNICAL SKILLS:**

**LANGUAGES/SCRIPTS**: Python (Tensor flow, NLTK, Network X, Prophet, Matplotlib, Pandas, Flask, PyTorch, SpaCy, cx, oracle), Java

**DATABASES**: MYSQL, PL SQL, Oracle, AWS RDS, AWS Redis, PostgreSQL, MongoDB

**TOOLS**: Hadoop, IntelliJ, VS Code, Anaconda, Jupiter Notebook, Jira, Apache Kafka, Apache Airflow, Jenkins, Docker, Tableau, AWS (S3, EC2, Elasticsearch, DynamoDB, SQS, Lambda, CloudWatch, Kibana, Sage Maker, RedShift, Quick Sight, Amazon EMR, Kinesis)

**CERTIFICATIONS**: AWS Certified Developer Associate

**PROFESSIONAL EXPERIENCE:**

**Amazon Fulfillment Technologies, Amazon**   **June 2022 – Till Date**

**Data Scientist**

**Responsibilities:**

* Part of a high functioning team that worked on HR solutions (Coaching, Performance feedback & scheduling) for Amazon Warehouses serving millions of customers across the world.
* Designed & Implemented Backend APIs of a tool that creates & stores feedback, based on the employees Productivity & Quantity, used in warehouses across world.
* Automated and scheduled computations that figure Feedback eligibility among warehouse employees which saved 3 hours of manual work/month across all warehouses globally.
* Implementing complex data manipulation logic using PL/SQL.
* Coordinate with Engineering team to design, validate, evaluate model deployment methodologies.
* Other responsibilities included communicating with Stakeholders, being a scrum master, solving customer tickets, backend or frontend software development, Unit & Integration testing, code deployment and bug fixes.
* Ensure that the testing team has a clear understanding of the expected behavior of the application.
* Develop a comprehensive test plan that outlines the scope, objectives, resources, schedule, and approach for functional testing.
* Define test cases based on functional specifications and user requirements.
* Build predictive models to optimize inventory management to reduce carrying costs while ensuring product availability by forecasting product demand, considering historical sales data, seasonality, and external factors.
* Part of a team that provide Optimized delivery routes to reduce shipping costs, reduce carbon footprint, delivery times, while adhering to driver & state policies.
* Developed end-to-end machine learning models, ETLs, Dashboards and scaled them to run in serverless production environments using AWS cloud services.

**Environment Tools** – Python (Beautiful Soup, NLTK, SpaCy, Scikit-learn, Flask, PyTorch, Pandas, NumPy, Matplotlib, seaborn, plotly), SQL, Java, PL SQL, React, Kotlin, Apache Kafka, NAWS tools (S3, EC2, Elasticsearch, DynamoDB, SQS, Lambda, CloudWatch, Kibana, Sage Maker, RedShift, Quick Sight, Amazon EMR, Kinesis)

**ASU Decision Theatre for Educational Excellence, Tempe, AZ**  **MARCH 2020 – June 2022**

**SR.DATA SCIENTIST**

**Responsibilities:**

* Leading a team that built multiple visualizations for high schools to visualize & assess their performance and provide feedback on how their course structure is translating to college enrollment.
* Built dashboards to Analyze high school student transcripts and performance data before and during the COVID-19 pandemic to assess the impact of disruptions caused by the pandemic on university enrollment.
* Develop data processing pipelines for large volumes of data using Hadoop and GCP - Set up a Hadoop cluster using Google Dataproc, design ETL processes to process large data sets & Load results into Big Query for analytics.
* Build a real-time data streaming pipeline for processing and analyzing data from multiple sources in real-time. Used PySpark for transformations, Google cloud storage, Big Query, automated this using Cloud Dataflow and scheduled with Cloud Composer.
* Migrated all the cleaned, transformed & encoded 1000+ school course tables from Oracle to GCP Cloud Datastore
* Built an ensemble of Ensemble ML models to label the school course data, to evaluate & automate student’s eligibility criterion for University Admissions office.
* Collaborate with database designers to ensure that the database schema is designed to support efficient PL/SQL development.
* Engineered ETL pipelines to analyze data from multiple sources using GCP Cloud Dataflow, enumerating the school resources, factors influencing the graduation rates, population demographics, community factors and funding in Arizona.
* Built Dashboards using Google Data studio, that visualize shifts in course selections, extracurricular activities, and student-led initiatives to evaluate the effects on BLM in schools & colleges.
* Develop various ML models to identify students who may require additional support for successful remote learning, considering factors such as access to technology and home environments.
* Create a dashboard that visualizes changes in student performance, attendance, and grades over time, highlighting specific courses or demographics that were most affected.

**Environment Tools** – Python (cx oracle, Beautiful Soup, NLTK, Apache Kafka, SpaCy, Scikit-learn, Flask, PL SQL, PyTorch, Pandas, NumPy, Matplotlib, seaborn, plotly), GCP – GCP Cloud Datastore, Big Query, Cloud Storage, Data prep, Cloud Composer, Data Studio.

**AGORDIA, AZ MAY 2018 – MAR 2020**

**DATA SCIENTIST**

**Responsibilities:**

* Only Data Scientist among a team of energetic & fast-paced developers, aiming to revolutionize the US Healthcare Realm.
* Designed multiple pipelines to filter, clean and process large data sets(~200Gb), to come up with valuable insights, metrics, valuable data aggregations and product for the company.
* Designed, Implemented and Deployed Agordia Store score, a metric used to rank around ~1 million pages in Agordia website.
* Participate in the development of data models and database structures.
* Develop and execute unit tests to ensure the reliability and correctness of PL/SQL code.
* Directed the build, storage, optimization, analysis and the visualization of Agordia’s proprietary US Healthcare Professional Social Interaction graph, with around ~800,000 nodes and ~40 million edges, to give us novel insights and growth strategies.
* Built custom NLP models (Sentimental Analysis, Event Extraction, Key-Word extraction, Synonym Nets) to analyze and leverage the medical channels in the app. These models have increased the task Throughput 10-fold, then manual labor.
* Designed, implemented & Deployed novel Identity Authentication API to protect ~2 million physicians & Agordia from HIPAA.
* Moderated support for multiple Growth Strategy pipelines and campaigns from a Social Network Analysis point of view. E.g.: using the social graph, figured who/to which specialty, a given Physician is strongly connected too.
* Handled the CI/CD pipelining for rapid releasing using Jenkins, Apache2, Docker Images, AWS EC2 and Selenium test suites.

**Environment Tools** –Python (Network X, SpaCy, TensorFlow, Flask, PyTorch, NLTK, Stanford NLP, Matplotlib, Seaborn, Pandas, NumPy, Beautiful Soup, selenium, scikit-learn, py-mongo), Neo4j, Gephi, VSCode, Tableau, AWS EC2, S3 & Sage Maker, Apache Kafka, PL SQL, MongoDB, Jupiter notebook, Confluence, Docker, CI/CD with Jenkins & Gi Kraken

**Genzeon May 2016 -Apr 2018**

**Data Scientist Co-Op**

**Responsibilities:**

* Participated in all phases of research including data collection, data cleaning, data mining, developing models and visualizations.
* Collaborated with data engineers and operation team to collect data from internal system to fit the analytical requirements.
* Redefined many attributes and relationships and cleansed unwanted tables/columns using SQL queries.
* Utilized Spark SQL API in PySpark to extract and load data and perform SQL queries and also used C# connector to perform SQL queries by creating and connecting to SQL engine.
* Performed data imputation using Scikit-learn package in Python.
* Performed data processing using Python libraries like NumPy and Pandas.
* Worked with data analysis using ggplot2 library in R to build visualizations for better understanding of customer behaviors.
* Implemented statistical modeling with XGBoost machine learning software package using R to determine the predicted probabilities of each model.
* Delivered the results with operation team for better decisions.

**Indian Institute of Technology, Hyderabad, INDIA AUG 2014 – APR 2016**

**RESEARCH Scientist**

**Responsibilities:**

* Developed & migrated the entire Twitter API from Java to Python using Tweepy library to facilitate creation of NLP models
* Provided support for OAuth & various queries – Timeframe, UserId, Geolocation & Keyword and stored them in Oracle.
* Implemented various Text Mining algorithms, Sentimental Analysis and various NLP classification algorithms using Python
* Opioid abuse detection through Twitter, an effort to curb Opioid abuse (Model Accuracy – 84.8%)
* Developed & deployed APIs for Ensemble models - SVM, NB and LSTM with word2vec to predict Opioid abuse.

**Environment tools**: Python (tweepy, cx oracle, sklearn, nltk, pandas, re, string, NumPy, Json, TensorFlow), Gi Kraken,

**DRDO, Bangalore, India Dec 2012 – May 2013**

**RESEARCH SCIENTIST Co-Op**

**Responsibilities:**

* Built a Prescription database, which is still in use to handle the medical records and transactions of the company.
* Technology used – Java, SQL, Tomcat, PhP, HTML, CSS, Microsoft SQL Server