SURYA TEJA SAITHANA

Chesterfield, MO (Open to relocation) • suryasaith@gmail.com • +1 (660) 238 8289 • LinkedIn • GitHub

PROFESSIONAL EXPERIENCE

BRIDGETOWN CONSULTING GROUP — **Piscataway, NJ**
Data Scientist — **Jan 2024 – Present**

- Contributed to training a **large language model (LLM)** using **Azure OpenAI Service** to autonomously write **Python** code for **EDA**. Employed **reinforcement learning with human feedback (RLHF)** and iterative refinement to enhance the model's performance. Achieved an **80%** improvement in the model's ability to perform coding and **business intelligence (BI)** tasks, revolutionizing AI model training and problem-solving methodologies.
- Developed a **price recommendation model** using existing annotated data, leveraging **ML** and **Python** libraries **(NumPy, Pandas, Scikit-learn)** on **Azure Machine Learning**. The model categorized customers into distinct classes, resulting in a **25%** improvement in pricing accuracy. Streamlined **ETL** processes using **SQL**, integrating diverse data sources such as **EHRs**, claims, and lab results, which facilitated seamless algorithmic development and deployment.
- Developed an end-to-end predictive analytics model on **Azure ML studio** to forecast patient readmission rates. The model effectively identified high-risk patients, allowing for targeted interventions that reduced readmission rates by **20%.**
- Engineered advanced machine learning models using Python's robust libraries to enhance customer classification, resulting in a **15%** increase in accuracy metrics. Integrated advanced deep learning models into production systems using **TensorFlow** and **PyTorch** on **Azure**, improving operational efficiency and predictive accuracy, significantly driving business outcomes through enhanced decision-making capabilities.
- Collaborated within interdisciplinary teams to ideate and implement data-centric solutions using big data and cloud technologies **(Hadoop, Databricks PySpark).** Effectively addressed business challenges by leveraging insights for strategic advantage, optimizing data processing and analysis workflows on Azure.

EMFOI INC — **Herndon, VA**
Data Science Intern — **Aug 2023 – Dec 2023**

- Conducted sentiment analysis within **Python** using **NLTK** and **SpaCy** library on customer product reviews and external websites through web scraping techniques.
- Contributed across the entire data lifecycle from acquisition to visualization, ensuring the delivery of comprehensive data science solutions.
- Leveraged Python libraries including **Pandas**, **NumPy**, and **Scikit-learn** to construct diverse machine learning models such as **Random Forest and stepwise regression**.
- Employed **cross-validation methodologies** to rigorously test models across various data batches, optimizing performance and mitigating **overfitting** risks.

QUALMINDS TECHNOLOGIES — **Hyderabad, India**
Data Scientist — **Jun 2020 – Aug 2022**
Clinical Trials Optimization, Cell Volume Prediction and Suspicious Activity Monitoring

- Developed and deployed **survival prediction and quality assessment models**, using **multi-modal analysis** for viability and titer, **leading to a 50%** reduction in laboratory experimentation duration. Leveraged these models to analyze historical experimental data and identify patterns in media compositions and cell survival, offering optimized nutrient, growth factor, and other component combinations to laboratory scientists.
- Created a cell culture analytics product using **R-Shiny**, incorporating a range of essential vitamins and amino acids. Enabled scientists to **predict cell survival rates** and formulate optimal media combinations, enhancing experimental accuracy and efficiency.
- Implemented rigorous data pre-processing techniques, including null value removal, **categorical feature encoding**, **data normalization, outlier removal, and data scaling**. Prepared cell culture data for analysis, ensuring high-quality input for downstream analytics.
- Designed robust data pipelines that seamlessly integrated data from various sources including cell culture development and cell retention devices, assisting in the enhancement of clinical trials.
- Transformed data management processes by translating **SAS** scripts to Python, developing **Apache Airflow and NiFi pipelines**, and optimizing **SQL** queries, achieving up to **20%** improvement in data processing efficiency and reducing query execution times by 20%.
- Enhanced data analysis and visualization capabilities using **Tableau** and **Power BI**, extracting critical insights from experimental data, and developed a **predictive analytics** product in **Python** for cell culture survival rates, leading to a **25%** improvement in experimental outcomes; led cross-functional collaborations with **DevOps** for seamless integration of data pipelines and ML models, ensuring operational efficiency and continuous model performance assessment for 10% increase in model reliability.
- Implemented **NLP** techniques **(SpaCy, NLTK)** in **Named Entity Recognition (NER)** for **Suspicious Activity Reporting**, improving entity detection accuracy by 35%. Trained a **Named Entity Recognition (NER)** model using the **SpaCy** package, achieving precise real time data processing through successful classification of NOUNS in test data sentences into predefined classes. Conducted experiments to determine the optimal number of clusters for trading data, integrating unsupervised algorithms into the production system and reducing the overall number of categories through effective categorization.
- Implemented advanced data modeling and analysis techniques with **Spark, SQL, and Talend** for data transformation/validation, reducing errors by **30%** and enhancing query efficiency, which supported strategic decision-making; directed database migration projects to new **SQL** Server environments, including schema changes and data migration coordination, while conducting regular database performance tuning, optimizing system performance by **25%**, and ensuring zero data loss during transitions.

IZIPPIE LABS — **Bengaluru, India**
Data Analyst/Data Scientist — **Jun 2016 – Jul 2017**

- Spearheaded the creation of the 'Resume Parser' system to analyze and categorize candidate profiles for Internal Job Postings (IJPs).
- Developed an algorithm using **pattern matching, regular expressions, and rule-based methods** to extract key resume components such as skills and experience.
- Implemented a relevance ranking system with machine learning techniques (**logistic regression, SVMs, ensemble models**) to rank candidates based on skills, experience, and job criteria.
- Streamlined recruitment by delivering top candidate lists to department managers and established rigorous validation procedures with cross-validation, manual verification, and metrics like precision, recall, and **F1-score.**
- Engineered and optimized **SSIS** packages for efficient data export tasks from **MS Access to SQL Server**, enhancing database performance and reliability.
- Implemented advanced data profiling techniques for data cleansing and standardization, significantly improving data quality by **20%**.
- Built **SQL** Server objects, including **Stored Procedures**, Views, and **User Defined Functions**, streamlining database operations and enhancing query performance.
- Optimized **SQL** Server databases by normalizing and indexing tables, and applying standard naming conventions, which reduced data retrieval times by **15%.**

SAPIENZA UNIVERSITY OF ROME — **Italy**
Graduate Assistant (Data Science) — **Oct 2017 – Jul 2020**

- Engineered a Python solution with machine learning (**K-Means, Gaussian Mixture Model**) for Vortical structure detection, increasing detection accuracy by **20%**.
- Employed **Principal Component Analysis (PCA)** for feature reduction and optimization, leading to a **30%** improvement in model efficiency.
- Utilized **visualization techniques (Arrow plots, Elbow plots)** for optimal **feature selection**, enhancing model interpretability and feature relevance.
- Developed and evaluated various clustering techniques for Vortical structure detection, achieving a **15%** enhancement in clustering precision.
- Created various models using **Regression, Support Vector Machines** for **predicting mechanical failures** in steam turbines, improving prediction accuracy by **25%.**

PROJECTS

Gender Prediction Model for Retail Customer Insights

- Developed **a predictive model** to determine customer gender based on purchasing behavior, achieving **92%** accuracy. The solution involved thorough **data cleaning, feature engineering, and model evaluation**, with a focus on **modular coding** for real-world implementation in a retail company's marketing strategy. (GitHub Link)

DIABETES SEVERITY DETECTION

- Employed **pre-trained CNNs** to identify stages of diabetes from **MRI scans** of the eye, achieving an **F1-score of 89%** and **8% improvement** over existing production models.

OPERATION RESOURCES MANAGEMENT

- Developed an inventory and resource management model using **Simplex** and **GRG Non-Linear** methods to optimize the cost function**,** enhancing operational efficiency for industries without **ERP software.**

EDUCATION

UNIVERSITY OF CENTRAL MISSOURI Warrensburg, MO, USA - ***Master of Science in Technology (Specialized in Data Science and ML)*** — **Aug 2022 – Dec 2023**
SAPIENZA UNIVERSITY OF ROME Rome, Italy - ***Master of Science in Engineering*** — **Aug 2017 – Jul 2020**
SRM UNIVERSITY Kattankulathur, India - ***Bachelor of Technology*** — **Jun 2012 – May 2016**

SKILLS

Programming, Databases, Data Visualization & Tools: Python, R, SQL, SAS, MongoDB, Jupyter Notebook, RStudio, Flask, Tableau, Power BI
Frameworks: Pandas, NumPy, Pytorch, TensorFlow 2.0, Scikit-learn, Seaborn, Matplotlib, Keras, OpenCV, SciPy, Spacy, NLTK, Joblib, xgboost, catboost,
Deployment Tools: GitLab, Jenkins, Dockers, Kubernetes, MLFlow
Machine Learning, NLP & Deep Learning: Supervised/Unsupervised Learning, Linear/Logistic Regression, Time Series, Classification, Clustering, Association Rules, Recommendation Systems, Regularization, Naive Bayes, Decision Trees, SVM, Ensemble Learning (XGBoost, ADABoost, Stacking), Hyperparameter tuning, Statistical Modeling,, Genetic Algorithm, Route Optimization, Gradient Descent, Text Preprocessing, NER, Regular Expressions, Lemmatization, Stemming, Topic Modelling, NLTK, Text Classification, SpaCy, Sentiment Analysis, Embeddings, Transformers, BERT, CNN, LSTM, Auto Encoders, Deep Neural Networks, Object Detection, Computer Vision (OpenCV)
Big Data & Cloud: Hadoop, Distributed Systems, Databricks PySpark, Hive, Azure ML Studio, Azure Databricks, AWS EMR, AWS Glue, Amazon QuickSight, Amazon SageMaker

CERTIFICATIONS

- **Microsoft Certified: Azure Data Scientist Associate, Credential ID:** 8487C96682307396**, Verification**
- **AWS Certified Machine Learning - Specialty, Credential ID:** 2b94d11694d34f729e8f343cfb5fe395**, Verification**