Introduction

As a highly accomplished Data Scientist with dual Master's degrees and extensive experience in machine learning, NLP, Gen AI, and predictive analytics, I am excited to share my portfolio with you. With a strong educational background and a proven track record of delivering impactful projects, I am confident in my ability to drive innovation and operational improvements across diverse industries. My passion for exploring AI and data science has led me to develop a unique blend of technical and business acumen, enabling me to tackle complex problems and deliver actionable insights.

I have a deep love for advanced machine learning techniques and data-centric solutions. I'm currently learning to leverage cutting-edge large language models (LLMs) and generative AI (Gen AI) to solve complex real-world problems. I aim to use these technologies to tackle challenges and positive impact. I firmly believe that learning is a continuous process in life. With this mindset, I'm always eager to expand my knowledge and skills. As I grow, I recognize that wisdom comes not just from accumulating knowledge, but from applying and reflecting on it to make a meaningful impact. This journey of learning and wisdom fuels my passion for tackling new challenges and continuously improving in the evolving field of AI and data science.

Education

Master's in Technology/Data Science, University of Central Missouri (August 2022 - December 2023)

Coursework: Machine Learning, Data Mining, Statistical Modeling, and Data Visualization

Thesis: "Developing a Predictive Model for Customer Churn using Machine Learning Techniques"

This program helped me develop a strong foundation in data science and machine learning, with a focus on practical applications and real-world problem-solving.

Master's in Mechanical Engineering, Sapienza University of Rome (September 2017 - June 2020)

Coursework: Control Systems, Thermodynamics, and Computational Fluid Dynamics

Thesis: STUDY OF VORTICAL STRUCTURES IN DNS DATA OF "INTERACTION BETWEEN AN OBLIQUE LAMINAR JET AND A PARTICLE□LADEN CROSSFLOW" USING ML

Engineered a robust Python solution for detecting Vortical structures within DNS data. This was based on the interaction between an oblique laminar jet and a particle-laden crossflow, employing machine learning algorithms like K-Means and Gaussian Mixture Model (GMM).

• Used the generated models to predict potential mechanical failures in the film cooling chambers of steam turbines, aiding in proactive maintenance and minimizing downtime.

• Led extensive data cleaning and Exploratory Data Analysis (EDA) initiatives, visually dissecting crucial data features and characteristics to inform machine learning models.

• Applied Principal Component Analysis (PCA) to the data, successfully reducing the feature set from over 20 to 15 without significant information loss. This step enhanced processing efficiency without compromising the data's integrity.

• Utilized visualization techniques, including Arrow plots and Elbow plots, to examine the cumulative

variance against the number of features, enabling optimal feature selection that retained 90% of the data variance.

• Implemented and evaluated various clustering techniques for Vortical structure detection. Assessed the performance by visualizing clustered data via scatter and heat plots, focusing on key features like vorticity and velocity.

Bachelor's of Technology in Mechanical Engineering, SRM University (July 2012 - May 2016)

Coursework: Mechanical Systems, Thermodynamics, and Materials Science

Professional Experience

- **Data Scientist, Bridgetown Consulting (January 2024 - Present)**
    - Developed and deployed advanced machine learning models to improve pricing accuracy and patient readmission rates
    - Collaborated with cross-functional teams to address business challenges and leverage big data and cloud technologies
    - Skills used: Python, Azure ML Studio, TensorFlow, PyTorch
    - Achievements:
        - Improved pricing accuracy by 25% using machine learning models
        - Reduced patient readmission rates by 20% using predictive analytics
        - Developed and deployed scalable machine learning models using Azure ML Studio
- **Data Science Intern, EMFOI INC (August 2023 - December 2023)**
    - Contributed to the development of comprehensive data science solutions across the entire data lifecycle
    - Employed cross-validation methodologies to enhance model performance and mitigate overfitting risks
    - Skills used: Python, Pandas, NumPy, Scikit-learn
    - Achievements:
        - Developed and deployed machine learning models to improve customer segmentation and targeting
        - Improved model performance by 15% using cross-validation methodologies

        - Collaborated with cross-functional teams to develop and deploy data science solutions
- **Data Scientist, Qualminds Technologies (June 2020 - August 2022)**
    - Developed and deployed survival prediction and quality assessment models using multi-modal analysis
    - Created a cell culture analytics product in R-Shiny to predict cell survival rates
    - Skills used: Python, R-Shiny, Pandas, NumPy
    - Achievements:
        - Improved model accuracy by 12% using multi-modal analysis
        - Developed and deployed a scalable cell culture analytics product
        - Collaborated with cross-functional teams to develop and deploy data science solutions
- **Data Science Graduate Assistant, Sapienza University of Rome (October 2017 - June 2020)**
    - Engineered a Python solution with machine learning algorithms to increase detection accuracy in vortical structure detection
    - Improved model efficiency through feature reduction and optimization using Principal Component Analysis (PCA)
    - Skills used: Python, Pandas, NumPy, Scikit-learn
    - Achievements:
        - Improved detection accuracy by 18% using machine learning algorithms
        - Developed and deployed a scalable Python solution
        - Collaborated with cross-functional teams to develop and deploy data science solutions
- **Data Scientist/ Data Analyst, IZippie Labs (June 2016 - July 2017)**
    - Spearheaded the creation of the 'Resume Parser' system using Python, Pandas, and NumPy
    - Implemented logistic regression, SVMs, and ensemble models for a machine learning relevance ranking system
    - Skills used: Python, Pandas, NumPy, Scikit-learn
    - Achievements:
        - Improved resume parsing accuracy by 22% using machine learning models
        - Developed and deployed a scalable relevance ranking system
        - Collaborated with cross-functional teams to develop and deploy data science solutions
    - Additional responsibilities:
        - Developed and maintained databases for data storage and retrieval
        - Worked with stakeholders to understand business requirements and develop data-driven solutions

Projects

Diabetes Severity Detection

- Problem Statement: Develop a model to identify stages of diabetes from MRI scans of the eye.
- Methodology: Employed pre-trained CNNs using Python and data science libraries such as Tensorflow, Pandas and NumPy
- Skills gained: Deep learning, computer vision, image processing
- Key Takeaways:
- Achieved an F1-score of 89% and 8% improvement over existing models
- Developed a scalable solution using transfer learning and fine-tuning of pre-trained models
- Improved model performance by using data augmentation techniques

Q&A ChatBot

- Problem Statement: Develop a Q&A ChatBot using Python, Streamlit, and Langchain
- Methodology: Leveraged the OpenAI GPT-3.5-turbo-instruct model for accurate and context-aware responses
- Skills gained: NLP, chatbot development, conversational AI
- Key Takeaways:
- Enhanced user engagement by 25% and achieved scalable software design
- Developed a conversational AI chatbot using transfer learning and fine-tuning of pre-trained models
- Link: [https://github.com/myaijournal/chatbot1]

Gender Prediction Model for Retail Customer Insights

- Problem Statement: Develop a predictive model to determine customer gender based on purchasing behavior
- Methodology: Performed data cleaning, feature engineering, and model evaluation using Python, Pandas, and NumPy
- Skills gained: Machine learning, data preprocessing, feature engineering
- Key Takeaways: Achieved 92% accuracy and developed a scalable solution with Object Oriented Programming concepts
- Link: [https://github.com/myaijournal/clientgender]

Sentiment Analysis of Corrona Virus Tweets

Link: [https://github.com/myaijournal/sentiment_analysis_corona_virus_tweets]

In collaboration with EMFOI Inc., the data science team embarked on a comprehensive Natural Language Processing (NLP) project focusing on sentiment analysis of tweets related to the Coronavirus. The project aims to develop robust machine learning models, uncover insights through topic modeling, and enhance our understanding of public sentiment during the pandemic. Project Overview: Our team, consisting of a lead data scientist and a senior data scientist, was assigned tasks related to data preprocessing, feature engineering, model development, and topic modeling. The primary dataset

comprises tweets containing information about the Coronavirus, with a sentiment label for each tweet. Accomplishments: 1. Data Preprocessing: - Concatenation: Merged relevant columns (`Location`, `TweetAt`, `OriginalTweet`) into a new column named `Tweet_texts` to streamline text analysis. - Text Cleaning: Employed regular expressions to remove special characters, dates, hyperlinks, hashtags, and usernames, ensuring the integrity of the text data. - Part-of-Speech Tagging: Utilized spaCy for part-of-speech tagging, enhancing our understanding of the grammatical structure and semantics of the tweets. 2. Exploratory Data Analysis (EDA): - Dependency Parser Visualization: Employed spaCy to visualize the dependency parser, revealing syntactic relationships within the text. - Named Entities Recognition: Leveraged spaCy for named entities recognition, identifying geolocation, money, and quantity entities within the tweets. - Visualization: Utilized spaCy's capabilities to visually represent the recognized entities in the tweets. 3. TF-IDF Analysis: - TfidfVectorizer: Utilized scikit-learn's `TfidfVectorizer` to transform the tweet texts into numerical representations using TF-IDF. - Cosine Similarity: Calculated the cosine similarity between the 200th and 20,000th tweets, providing a measure of their textual resemblance. - Corpus Vector: Computed the corpus vector as the average of all document vectors, offering a representative vector for the entire dataset. 4. Sentiment Analysis Models: - Random Forest Model: Developed a sentiment analysis model using a random forest classifier within a scikit-learn pipeline, facilitating seamless data processing and model training. - Evaluation: Assessed the model's performance on the test dataset, considering metrics such as accuracy, precision, recall, and F1 score. Justified the model's effectiveness based on these evaluation metrics. 5. Advanced Sentiment Analysis Models: - Hyperparameter Tuning: Applied grid search cross-validation to fine-tune hyperparameters for both the TfidfVectorizer and the random forest classifier in a pipeline. - Model Diagnosis: Evaluated the performance of the tuned model on the test dataset, providing a detailed diagnostic analysis to justify the model's effectiveness. 6. Topic Modeling: - LDA and CountVectorizer: Implemented Latent Dirichlet Allocation (LDA) for topic modeling and used CountVectorizer to convert text data into a document-term matrix. - Visualization: Presented visualizations showcasing the top 15 words associated with each of the five topics derived from the first LDA model. 7. Advanced Topic Modeling: - LDA with TfidfVectorizer: Extended topic modeling using LDA with TfidfVectorizer, considering the importance of terms within the context of the entire corpus. - Visualization: Illustrated the top 15 words for each of the five topics in the second LDA model. - Dimension Reduction: Applied dimension reduction techniques to visualize the second topic model, offering a concise representation of the topic relationships.

Portfolio ChatBot: A Vector-Search Powered Conversational AI

GitHub Link: [https://github.com/myaijournal/portfolio_chatbot]

HuggingFace Spaces Link : [https://huggingface.co/spaces/myaijournal/portfolio_chatbot]

Portfolio ChatBot is a cutting-edge conversational AI project that leverages vector search and natural language processing to provide users with accurate and informative responses. The project consists of six key components:

1. Database Creation: Astra DB is used to store and manage the vector database.
2. PDF Processing: Text is extracted from a collection of PDF files using pdfminer.six.
3. Vector Database: The extracted text is embedded and loaded into Astra DB using the Python client.
4. Chatbot Development: A chatbot is built using LLaMA 3.1 and LangChain, featuring a prompt template that matches user queries to the vector database.
5. Web Development: A website is created using Streamlit/Django to provide a user-friendly interface for interacting with the chatbot.
6. Deployment: The website is deployed using Hugging Face Spaces.

Skills Involved:

- Database management (Astra DB)
- PDF text extraction (pdfminer.six)
- Vector search and embeddings (Astra DB)
- Chatbot development (LLaMA 3.1, LangChain)
- Web development (Streamlit/Django)
- Deployment (Hugging Face Spaces)

Key Takeaways:

- Building a conversational AI that leverages vector search for accurate responses
- Integrating PDF text extraction and vector embeddings into a chatbot
- Creating a user-friendly website using Streamlit/Django
- Deploying a project using Hugging Face Spaces
- Combining multiple technologies to build a cutting-edge conversational AI

Technical Skills

As a seasoned data scientist, I possess a broad range of technical skills that enable me to tackle complex projects and deliver impactful solutions. My technical expertise spans programming languages, databases, data visualization tools, frameworks, deployment tools, machine learning, NLP, deep learning, Gen AI, big data, and cloud computing.

Programming Languages

- Python
- R
- SQL
- SAS

Databases

- MongoDB
- Cassandra DB
- Astra DB

Data Visualization & Tools

- Jupyter Notebook
- RStudio
- Tableau
- Power BI

Frameworks

- Pandas
- NumPy
- Pytorch
- TensorFlow 2.0
- Scikit-learn
- Seaborn
- Matplotlib
- Keras
- OpenCV
- SciPy
- Spacy
- NLTK
- Joblib

- XGBoost
- Catboost
- Flask
- Streamlit
- PyPDF2

Deployment Tools

- GitLab
- Jenkins
- Dockers
- Kubernetes
- MLFlow

Machine Learning, NLP & Deep Learning

- Supervised/Unsupervised Learning
- Regression
- Time Series
- Classification
- Clustering
- Recommendation Systems
- Regularization
- Naive Bayes
- Decision Trees
- SVM
- Ensemble Learning (XGBoost, ADABoost, Stacking)
- Hyperparameter tuning
- Statistical Modeling
- Gradient Descent
- Text Preprocessing
- NER

- Regular Expressions
- Lemmatization
- Stemming
- NLTK
- Text Classification
- Sentiment Analysis
- Embeddings
- BERT
- CNN
- LSTM
- Auto Encoders
- Deep Neural Networks
- Computer Vision (OpenCV)

Gen AI

- LLMs
- Llama
- LangChain
- HuggingFace
- OpenAI
- Vertex AI
- Prompt Templates
- Sequential Chains
- Chat Models
- Transformers
- Output Parsers
- Azure OpenAI Service

Big Data & Cloud

- Hadoop

- Distributed Systems
- Databricks PySpark
- Hive
- Azure ML Studio
- Azure Databricks
- AWS EMR
- AWS Glue
- Amazon QuickSight
- Amazon SageMaker

I am confident that my technical skills, combined with my passion for data science and machine learning, enable me to deliver high-quality solutions that drive business impact.

Certifications

I have obtained several certifications that demonstrate my expertise and commitment to staying up-to-date with industry developments in data science, machine learning, and cloud computing. These certifications include:

- **Microsoft Certified: Azure Data Scientist Associate**
    - Demonstrates my ability to design and implement data science solutions using Azure Machine Learning, Azure Databricks, and other Azure services.
    - Link [https://learn.microsoft.com/en-us/users/suryatejasaithana-8678/credentials/8487c96682307396]
- **AWS Certified Machine Learning - Specialty**
    - Validates my skills in designing, implementing, and deploying machine learning models using AWS services such as Amazon SageMaker, AWS Glue, and Amazon QuickSight.
    - Link [https://cp.certmetrics.com/amazon/en/public/verify/credential/2b94d11694d34f729e8f343cfb5fe395]

Other Interests

In addition to my passion for data science and machine learning, I enjoy a variety of hobbies and interests that help me maintain a healthy work-life balance and foster creativity.

- **Sports:** I am an avid sports enthusiast, particularly enjoying Cricket and Badminton. Playing sports helps me develop teamwork, strategy, and problem-solving skills, which I also apply to my work in data science.

- **Traveling:** I love exploring new cultures and destinations. I have been fortunate enough to travel to several countries, including Italy, France, Croatia, and Greece. Traveling broadens my perspective, helps me understand diverse viewpoints, and inspires me to approach problems from unique angles.
- **Cooking:** I enjoy experimenting with new recipes and cuisines in my free time. Cooking helps me develop creativity, attention to detail, and patience – skills that also benefit my work in data science.

These interests help me recharge, expand my horizons, and bring fresh ideas to my work in data science. I believe that having a well-rounded life outside of work is essential for maintaining productivity, creativity, and overall well-being.

Contact Information:

I'm looking to collaborate on innovative AI and machine learning projects that push the boundaries of what's possible. Dive in, explore my projects, and feel free to reach out if you have any questions or just want to chat about all things AI and data science!

How to reach me: suryasaith@gmail.com | https://www.linkedin.com/in/suryasaithana/ | https://github.com/myaijournal