Raj De | PhD — Data Scientist

(315) 949-8690 | dedederaj@gmail.com | LinkedIn | GitHub

Data Scientist with a PhD in Physics and a proven track record in machine learning, and data-driven decision-making. Experienced in developing statistical models, analyzing large datasets, and delivering actionable insights.

Skills

PROGRAMMING LANGUAGES: Python | C | SQL | LATEX | HTML
VERSION CONTROL: GIT
MACHINE LEARNING LIBRARIES: Scikit-Learn | PySpark | Statsmodels | CatBoost | XGBoost
DATA PROCESSING & VISUALIZATION : PostgreSQL | Numpy | Pandas | GeoPandas | Pytorch | Matplotlib | Seaborn | Plotly | Flask | Streamlit
CLOUD COMPUTING & DATA ENGINEERING : AWS S3 Bucket | AWS Glue | Amazon SageMaker
INDUSTRY KNOWLEDGE: Scientific Computing| Supervised Learning | Unsupervised Learning | Quantitative Research | Statistical Analysis | Deep Learning | Natural Language Processing | Data Mining | Large Language Modeling

Education

Ph.D. in Physics | Syracuse University – May 2023

Data Scientist Fellow | The Data Incubator – March 2023

Python for Data Science AI & Development | IBM

M.S. in Physics | Syracuse University – June 2017

Experience

Research Scientist| Syracuse University | Syracuse, NY| Sept 2018 – December 2022

- Successfully conducted and completed research on four projects on geometric studies on thin films, with one of the projects featured in American Physics Society Magazine.
- Reduced simulation runtime by 40%+ by applying new approach in mesh models optimizing energy of floating thin films using gradient descent.
- Developed robust data ingestion pipeline, employing Python computing and modules to effectively manage and preprocess more than 600 GB of simulation data. Implemented strategic data transformation steps, ensuring seamless data processing and analysis.

Projects

1. Anomaly Detection

- Developed solutions for monitoring client's business metrics in real time for instant detection of incidents that may impact their revenue.
- Leveraged ensemble models and conducted comprehensive performance analysis utilizing precision, recall, and F1 score metrics. Identified an optimal model with a 15% increase in true positive identification, leading to enhanced anomaly detection capabilities.

2. Trails To Health

- Designed and implemented "Trails to Health," a recommendation system for trails across 230 New York State Parks.
- Built unsupervised learning model to cluster trails by difficulty, accounting for factors like elevation and distance. and employed cosine similarity for user recommendations.
- Integrated USGS Rest API, Flask, and geolocation services for streamlined user access and personalized trail recommendations. (Trails To Health).

3. Autumn Leaves

- Led a project to develop a fall leaf color prediction model for Vermont's Mount Mansfield, employing SARIMA modeling and time series analysis of data.
- Created a web application providing historical and forecasted leaf color changes, enhancing visitor experiences. (Ongoing)

Volunteer Work:

UN Datathon 2023

- Led team of 5 data scientists to tackle complex climate data challenge focusing on fire hotspot prediction in Sumatra.
- Utilized AWS S3 buckets, AWS Glue, and Amazon SageMaker for efficient data processing and integration of fire hotspot and climate data.
- Developed predictive ML models to enhance early warning systems for fire outbreaks, potentially adding in disaster prevention and mitigation.

DataKind September 2023 Event

- Collaborated with a diverse team of over 200 volunteers to support the IMF Gender-Based Project, aimed at promoting gender equality for sustainable economic progress across 5 countries.
- Spearheaded the development of a document translation pipeline, facilitating the conversion of non-English documents to English for seamless communication and data analysis.
- Leveraged Large Language Model OpenAI technologies through Azure cloud computing to create an interactive PDF insights chatbot providing users with a more interactive approach to exploring data and conducting in-depth analysis.