Raj De | PhD — Data Scientist

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

Data Scientist with a PhD in Physics and experience in applied data science. Skilled in extracting insights from complex datasets, with a proven track record in machine learning, statistical analysis, and data-driven decision making.

Skills

PROGRAMMING LANGUAGES: Python | C | SQL | LATEX | HTML
VERSION CONTROL: GIT
MACHINE LEARNING LIBRARIES: Scikit-Learn | PySpark | Statsmodels | CatBoost | XGBoost | Langchain
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| Error Analysis | Temporal Prediction | Statistical Analysis | Data Transformation | Deep Learning | Natural Language Processing | Data Mining | Large Language Modeling

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 SARIMAX modeling and time series analysis of data.
- Expanded the project's scope by implementing and evaluating various regressor models like Catboost, XGBoost, SVR and Random Forest Regressor, ultimately selecting Catboost since it outperformed other models by nearly 10 percent.
- Created a web application providing historical and forecasted leaf color changes, enhancing visitor experiences. (Ongoing)

4. AI Detects Peat-fires in Sumatra: Unmasking with Technology - **UN Datathon 2023**

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

5. Linking Household Survey Questions & Data - IMF Gender & Finance DataDive - **DataDive September 2023**

- 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 as a volunteer for **DataKind**.
- 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.

6. Daily Sales Problem - **Novartis Datathon 2023**

- Built a daily sales phasing forecast model for the Novartis Datathon 2023, focusing on a key metric that measures the ratio of sales per working day within a month to enhance the accuracy of monthly sales forecast.
- Conducted data exploration through bar plots to identify outliers and correlation plot to identify important features.
- Developed a LSTM model to forecast daily sales phasing for each country-brand combination, achieving a mean square error of 0.001.

- Trained a CatBoost ensemble model for 780 country-brand combinations, as less computationally expensive alternative for LSTM with mean square error of 0.001 on testing data demonstrating robust performance and generalization capability.

Experience

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

- 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 large scale datasets(>600 GB). Implemented strategic data transformation steps, ensuring seamless data processing and analysis.
- 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.
- Presented complex data findings to both technical and non-technical stakeholders, ensuring clarity and actionable insight.

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