ABHIJITH RAJAN

Data Scientist

abhijithrajan@gmail.com   linkedin.com/in/abhijith-rajan   github.com/abhijithrajan

Summary

Data scientist with over 10+ years of experience in research, data analysis, model design and simulations. Skilled in Python programming, machine learning, unstructured data, statistics and problem solving.

In the past year, I have worked as a Data Scientist at a startup in New Jersey named Mareana Inc. At Mareana I have worked on multiple projects with different clients, where I have been primarily involved in extracting insights from unstructured data, including images and/or text files, and also building forecast models for time-series data. As a member of a startup, I have worn several different hats, including Project Manager, Data Scientist/Engineer, and Data Analyst. I have built and delivered software tools that were available via an online API. I am also in charge of hiring of new Data Scientists and Analysts for our firm.

Prior to this, as an Astrophysicist, I gained significant experience working with Imaging data applying different techniques to extract small signals from noisy data. As an analyst, I led a project to build the first pipeline to image planets outside our solar system using the Hubble Space Telescope. I was also the project lead for a simulation tool to explore and characterize the utility of several new instruments for the next generation of space telescopes.

I am now looking for new opportunities for growth. My ideal role will allow me to improve my understanding of traditional machine learning methodologies while giving me the freedom to explore new technologies.

Skills

Languages	Python, SQL
Database	MySQL, PostgreSQL, neo4j
Statistical Methods	**Regression Methods:** Linear, Polynomial, Lasso & Ridge Regression **Classification Methods:** Logistic Regression, K-NN, Decision Trees. **Ensemble Learning:** Random Forest Models (Bagging & Boosting) **Text Mining:** Natural Language Processing, Sentiment Analysis. **Time Series:** , AR, MA, ARMA, ARIMA, ARCH & GARCH.
Visualization Tools	Matplotlib, Seaborn, Bokeh, PowerBI.
Cloud Services	AWS S3, EC2, Google Colab, Google Compute.

Professional Experience

Mareana Inc   **Princeton, NJ**
Data Scientist   **Jan 2019 - Present**

Description:

Headquartered in Princeton, NJ, Mareana helps companies solve complex data problems, including searching unstructured data and predictive analysis on this data. At Mareana, I have worked with several clients, listed below.

Johnson & Johnson   **Princeton, NJ**
Data Scientist   **Jan 2019 – June 2019**

Responsibilities:
- Johnson and Johnson are a long-time customer of Mareana's. The first project I was a part of was the SMARTFIND project where we were tasked to extract and present all the information contained in a large corpus of documents (~10000 – 50000).
    - For the SMARTFIND project, I built an automated pipeline to extract information contained within design documents.
    - Information in design documents included autolocating and extracting tabular information, specific parameters relevant to each file such as title, drawing number, etc and finally any text that might be present in the image.
- The second project was to collect information across several different plants and combine them into a single dashboard.
    - For the project, I built a natural language query parser that was made available via a RESTful API. The primary purpose of which was to allow regular users to query the data without knowing SQL.
    - The tool provided text autocomplete, query completion suggestions, and converted simple language queries to their SQL equivalent.

Technologies: Python, SQL, Flask, Swagger, Pandas, Trie, OpenCV, Tesseract.

Merck & Co.   **Branchburg, NJ**
Data Scientist   **Mar 2019 – Present**

The Merck Integrated Product Information (IPI) project is a multi-year effort to combine all information regarding Merck products globally into a single, structured, document management system.

Responsibilities:
- Collaborated to build a pipeline that extract all data present in different PDF documents into XML. Data typically includes text, tables, footnotes, mathematical formulae, etc. The XML data are further parsed into a SQL database.
- Documents are continually updated, so the pipeline tracks all modifications and maintains a repository that logs the changes.

Technologies: Python, SQL, Pandas, eTree, Adobe Acrobat API.

Roche Molecular Systems Inc   **Branchburg, NJ**
Project Manager / Data Scientist   **Apr 2019 – Present**

The team at Roche presented us with three years of historical data detailing the process involved to produce different items at a single plant. Mareana is currently building a forecasting tool to help planners at Roche.

Responsibilities:
- Performed Exploratory Data Analysis on ~4 million historical purchase orders to find appropriate distributions to help model the data.
- Performed decomposition to examine for trends and stationarity of data. Used time series forecasting methods such as exponential smoothing, and ARIMA.
- Modeling included providing best estimates of down-times or other issues, calculated using the historical data.
- Real time forecasts were made available through API calls.
- Also working with a separate business team to utilize historical information to help improve efficiency of operations on the shop floor.

Technologies: Python, Statsmodels, Pandas, Flask, Swagger.

Space Telescope Science Institute   **Baltimore, MD**
Research Fellow   **Jun 2017 – Jan 2019**

Description:
Space Telescope Science Institute (STScI) is one the world's premier Astrophysical research laboratories. STScI is responsible for managing the Hubble Space Telescope and the next generation James Webb Space Telescope. Researchers at STScI focus on advancing not just the field of Astrophysics but also instrumentation and data processing.

Responsibilities:
- Generated a Python-based simulation tool for use with specific coronagraphic instruments on ground and space-based instrumentation. Improved simulations of data from the James Webb Space Telescope to incorporate realistic noise and thermal models.
- Used least squares and PCA based techniques to transform and extract insights from imaging data.
- Modeled exoplanet spectra via Markov-Chain Monte Carlo methods to estimate the composition of exoplanet atmospheres.
- Mentored multiple Johns Hopkins University graduate students. Co-founded the Chesapeake Bay Area Exoplanet Meeting series.

Technologies: Python, Astropy, PyFits, pyKLIP.

Space Telescope Science Institute   **Baltimore, MD**
Research and Instrument Analyst II   **Sep 2008 – Apr 2012**

Responsibilities:
- Member of the Wide Field Camera 3 team on the Hubble Space Telescope. Led multiple programs to study the visible and near infrared camera performance, including long term monitoring of noise properties, and camera degradation due to cosmic radiation.

- Co-developed a method to image extra-solar planets for space telescopes. The method is now a standard observing mode for the next major NASA space telescope, the James Webb Space Telescope.
- Mentored undergraduate students participating in the Space Astronomy Summer Program.

Technologies: Python, Astropy, PyFits, Mathematica, IDL.

Other Activities
- Graduate of 'The Data Incubator' Fellowship. The program chooses a small group of PhDs from across the US to help enter the field of Data Science.
- Data Science mentor at BSA Spec, which is a local consultancy. I work with new consultants interested in becoming Data Scientists.

Education
Doctor of Philosophy, Astrophysics   Arizona State University | Tempe, AZ | May 2012 – May 2017
Master of Science, Astronomy   San Diego State University | San Diego, CA | Aug 2005 – Aug 2008
Bachelor of Engineering, Instrumentation University of Pune | Pune, India | Sep 1999 – Aug 2004