Rahul Mathur

Senior Data Science Manager / Principal Data Scientist

(+91) 9663619000 | rahul10mathur@gmail.com| LinkedIn

Summary

- Result driven professional with 14 years' work experience helping organizations leverage data driven decision making
- Rich experience of over 7 years in Data Science and prior experience in Data Analytics & Engineering domains
- Spearheading Machine Learning & Deep Learning initiatives across Marketing, Risk & Identity analytics
- Mentoring and managing a team of Data Scientists at all levels.
- Collaborating across teams - Product, Engg & DevOps. Conceptualizing and implementing models on distributed systems
- Hands on Individual with proficiency in Python, PyTorch (deep learning), Apache Spark, Hadoop and Cloud (AWS) ecosystems

Work Experience

Senior Data Science Manager – Neustar Inc.

May 2017 – Present
Bangalore, India

Currently working as Sr Data Science Manager at Neustar Inc., a US based product firm

Role & Responsibilities

- Working with Product team to conceptualize analytical visions into data science problems and drawing the value proposition.
- Driving large scale projects in ML & DL space working with petabytes of data.
- Deploying models in **distributed systems** in collaboration with Data Engineering & DevOps.
- Mentoring & managing Data Science team, interviewing and hiring candidates. Resource's training & upskilling plans.
- **Forming R&D capabilities** for exploring research areas towards enhancing products. Using **Agile methodology** for planning

Key Initiatives

Contextual Advertising (Marketing Analytics)

Marketing industry is getting disrupted due to removal of cookies, restriction in First Party data to name a few. Obtaining context (via NLP techniques) of the web page user is visiting & keywords could help segment the audience and aid targeted marketing

- Building a distributed and generic web scraping framework using Scrapy to capture the static content of the web pages
- Performing **NLP** based data preprocessing pipeline. **Topic Modeling** using LDA to formulate the Categories under which the web pages are found.
- Improving Topic coherence using Neural Topic Modeling approaches – Prod LDA (an **Encoder-Decoder** based model) along with pretrained **BERT model**. Led to more interpretable Topic clusters
- Using IAB Taxonomy, achieved **83% accuracy**, match rates for the categories
- Distributed architecture using Docker, Kubernetes along with AWS cloud.

Risk Modeling

Neustar possesses rich and quality PII data along with IP risk data and vendor sourced Geo location data. These orthogonal signals provide key insights to detecting Fraudulent behavior and building Risk profile of Individuals.

- Built **Models in Fraud Analytics domain** by using signals from Geo Location, IP data, Personal Identifiable Information data
- Heavily worked on class imbalance techniques. Utilized boosting algorithms –**Light GBM** to get lifts in ROC curves & Precision Recall curves. Defined key metrics – Fraud Omission Rate , with high precision in top 3% population
- Also explored Multilayer Perceptron on raw feature vectors to further gauge improvements in precision metrics.
- Tech stack – Pytorch , Python, Scikit Learn ,Scipy, Kubernetes ,Docker , AWS ,Tableau ,Hive ,SQL

Entity Resolution - Identity Analytics

Core backbone of Neustar products is based on **People graph with offline and online attributes**. Data is sourced from various first & third-party providers and a key challenge was to build a state-of-the-art Entity Resolution framework that captures all inputs around US population (~ 330 M) in righteous clusters

- Developed algorithmic framework to resolve Person entity data for over **5 billion records**. Utilized Decision Trees Classifier & curated Clustering based approaches (DB Scan & Hierarchical) to generate high quality data clusters of Individuals

- Used First party & Third-party data with varied attributes to determine clusters of Individuals & Businesses. Directed to US population of nearly 330 M, curated an Identity resolution process **graph** with **F1 Score of 93%.** stitching Offline and Online based signals
- Tech Stack – Spark ML , PySpark , Scikit , HDFS , Python , Scala

Linkage Models – Identity Analytics

- **Conceptualized Linkage Models** using Logistic Regression, Random Forest to determine the most recent and accurate attributes linked to an individual clusters using external signal sources and deriving a timeline of associations
- Towards Offline signals of Email, Phone & Date of Birth linkages, model had **AUROC of 87%.**
- Developed a Business-Person Email names identification model using NLP & Classification approaches, with high precision
- Tech Stack used –SparkML , Spacy , POS Tagging , NER , Jupyter Hub, AWS EMR , HDFS ,Hive , Tableau

Senior Technology Consultant (Data Science)– PriceWaterhouseCoopers

July.2015 - May 2017
Washington D.C US /Bangalore, India

- Crafted an Anomaly Detection framework for a Log parsing solution built on Elasticsearch-Apache Spark- Kafka-Fluent data pipeline. Implemented as part of POC's to couple of clients
- Used IP data to determine diverse regions accounting for high volume of logs to predict more malicious cases
- Predicting GCP cost usage per account based on the infrastructure utilized which helped in RFP's for new accounts
- Text analytics-based model to determine user's sentiment in Claim's process for leading US Insurance giant. Used TF-IDF, Naïve Bayes approach
- Tech Stack – Python, Pandas, Scikit learn, NLTK , Google Cloud Platform (GCP) , Elasticsearch ,Kibana ,Airflow

Project Lead – Mindtree Ltd

Dec.2011 - July 2015
Bangalore, India

- Architected web scraping based framework for capturing all wells for Oil & Gas companies. It provided high risk zones and cross selling opportunities.
- Developed a de-duplication application for People identities using Similarity algorithms
- Tech Stack – Python, BeautifulSoup, Postgres DB, MySQL , HDFS, Hive, Sqoop , Qlikview reporting tool

Data Engineer – 24/7.ai

Jun 2010 - Dec 2011
Bangalore, India

- Implemented solutions in ETL pipeline, data preparation, databases set up, DDLs, Views. Built reporting dashboards
- Tech Stack – Perl, Pentaho , Shell scripting , MySQL

Software Engineer – Accenture

Jun.2007 - Jan 2010
Bangalore, India

- Backend engineer building data pipeline, batch operating jobs and reporting frameworks.
- Tech Stack – Shell scripting, Oracle DB , Actuate reporting tool

Others

Data Science Consultant, Freelance – Jigsaw Academy Institute (Analytics Education)

Dec 2015- May 2017

- Developed course on Machine Learning in Python. Content included Statistical modelling, Regression & Classification techniques along with coding exercises using real world datasets
- Work recommended (here) by COO of Jigsaw Academy
-

Education

Master's in Science (MSc) in Analytics *2017-2019*
(Birla Institute of Technology and Science, Pilani, India)

Executive Program in Business Analytics (EPBA) *2014-2015*
(Indian Institute of Management, Calcutta, India)

Bachelor's in Engineering in Information Science *2003–2007*
(Visvesvaraya Technological University (VTU), Karnataka, India)

Indian School Certificate (ISC) *2001-2002*

Indian Certificate of Secondary Education (ICSE) *1999-2000*

Certifications

AWS Certified Solutions Architect – Associate Level License AWS-ASA-31704
Mathematics for Machine Learning (Imperial College London)

Technical Proficiency

Scripting Languages	Python 3.x, PySpark , SQL ,
Frameworks	Scikit , NLTK , Spark ML, Pytorch (Deep Learning)
Machine Learning	Linear & Trees based Models, Dimensionality Reduction techniques (PCA , SVD) , A/B testing Ensemble Models & Boosting (XGBoost , LGBM)
Deep Learning	Pytorch, RNN, LSTM , GRU, Encoder- Decoder, Multilayer Perceptron
Cloud	Amazon Web Services (AWS), Google Cloud Platform (GCP)
Big Data Stack	Apache Spark 2.2, AWS EMR , HDFS, Hive, Spark SQL
Key-Value stores & RDBMS	MySQL, Elasticsearch, Redis
Visualization Tools	Tableau , Kibana
Open Source	Virtual Box, Github, Jira, Airflow, Scrapy (Python tool)
IDE	Jupyter, PyCharm
Processes	Agile Methodology ,Sprint, Scrum, Performance reviews

Personal Information

DOB	10th Feb 1984
Address	515 A, Esteem Enclave, Bannerghatta Road, Bangalore – 76 , India
Nationality	Indian
Visa	Posses valid B1 US Visa