SONAL JAIN

linkedin.com/in/sjain2212/ | jain.son@northeastern.edu | github.com/sonaljain2212 | Boston, MA | 8579305296

EDUCATION

Northeastern University, Boston, MA Expected Graduation: Aug 2021

Master of Science in Data Science (GPA- 3.67/4) Sep 2019 - Present

Courses: Supervised and Unsupervised Machine Learning, Causal Modeling, Natural Language Processing, Algorithms and Data Structures, Data Management and Processing, Deep Learning.

Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India Aug 2012 - Jun 2016

Bachelor of Engineering in Electronics and Communication Engineering (GPA- 3.6/4)

Courses: Linear Algebra, Statistics, Databases, Algorithms, Data Structures, Machine Learning.

TECHNICAL SKILLS

Programming Languages:	Python, R, SQL, Java
Data Science skills:	Data Cleaning, Data Analysis, Predictive Modeling, Data Visualization, NLP, Deep Learning, Google Cloud Platform, Hypothesis testing, Statistical Modeling, AWS, A/B testing
Libraries/framework:	Pandas, Numpy, , Scikit-learn, NLTK, spaCy, PyTorch, Dplyr, Tensorflow, PySpark, Flask, Gensim
Data Visualization:	Matplotlib, Seaborn, Plotly, Ggplot2, Tableau, PowerBI
Tools/Technologies:	RStudio, Jupyter Notebook, Advance Excel, Docker, Git, Github, MS Office, JIRA, Agile, SDLC

PROFESSIONAL EXPERIENCE

Quantiphi Inc, Data Science Consultant Intern, Marlborough, Massachusetts, US Aug 2020 – Dec 2020

- Effectively reduced manual efforts, time, and cost by **95**% by developing and improving ML pipelines for **DOCUMENT AI** web-app with **8** critical features to process documents using **Machine Learning** and **Google Cloud Platform APIs**.
- Included core features like **Classification**, **Entity Extraction**, **Purchase Order Processing** with **98%** classification accuracy.
- Incorporated **Document Translation**, **Summarization**, **Searchable Document**, **Healthcare Entity and Relationship Extraction**, **Driver's License Entity Extraction** using SOTA NLP models and presented demos with customized requirements to clients.

Tata Consultancy Services (TCS), **System Engineer**, Mumbai, Maharashtra, India Oct 2016 - July 2019

- Saved employee's time by creating **Course Recommendation System** suggesting **top 5** courses similar to employee's previous choice and collaborated with engineering and design team to incorporate this new feature in the portal.
- Incorporated feature reduction (**PCA)** and **Linear Regression** for rating prediction of new courses and unrated courses.
- Presented the analysis of the top areas, courses, and improvements to the higher management and stakeholders.
- Developed selenium automation testing framework and achieved **50%** reduction in testing time on **500** scripts daily.

ACADEMIC PROJECTS

Deep Image Prior (Python, CNN)

- Improved image quality by implementing randomly initialized Convolutional NN (SkipNet) as prior for **image denoising**, **super-resolution**, **inpainting** without training CNN and compared performance with trained SOTA models using **PSNR** as metric.

Healthcare Entity and Relationship Extraction on Clinical Data (Python, NLP)

- Expedited screening of patient records by extracting medical NERs and relationship between drugs, dosage, frequency, route using pre-trained **Med7** and fine-tuned **BioBERT** on N2C2 clinical patients' dataset and displayed using relation graphs.
- Conducted Market Basket and Cluster Analysis using **Apriori**, **DBSCAN, k-means, TSNE** to group and recommend similar NERs.

Universal Sentiment Analyzer Application (Python, NLP, Streamlit, Heroku) Link

- Simplified sentiment analysis process by **85%** by developing a multi-data Universal Sentimental Analysis App that performs text **Data Analysis, Tokenization, Vectorization** and **Modeling** using statistical ML models and LSTM on sentiment datasets.

Credit Card Fraud Prediction (Python)

- Analyzed important contributors to fraud transactions, performed feature engineering, predicted fraudulent transactions on highly imbalanced target class to achieve recall of **70 %** using 2-layer **Neural Network** with weighted loss function.

Customer Churn and Retention Prediction (Python)

- Recommended new metrics to measure retention and quality of service by analyzing Job board data, suggested measures to decrease customer churn and predicted high retention of **98%** recall using Random Forest and AdaBoost Classifier.

Return on Investment prediction using Causal modeling on Boston Airbnb data (Python, R, Pyro)

- Economized the buyers by creating a model that finds highest **ROI** properties and best areas to buy properties using **Causal Modeling** (probabilistic programming) and **Bayesian Statistics** techniques.