SONAL JAIN

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

EDUCATION

Northeastern University, Boston, MA Sep 2019 -Aug 2021
Master of Science in Data Science (GPA- 3.67/4)
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, MySQL, NoSQL, Java, MATLAB
Data Visualization:	Matplotlib, Seaborn, Plotly, Ggplot2, Tableau, PowerBI
Data Science skills:	Data Cleaning, ETL, Data Analysis, Predictive Modeling, Data Visualization, NLP, Deep Learning, CNN, GCP, Statistical Modeling, Time-Series Forecasting, Automation, AWS, A/B testing
Libraries/Framework:	Pandas, Numpy, Scikit-learn, NLTK, spaCy, Dplyr, PyTorch, Tensorflow, Apache Spark, Flask
Tools/Technologies:	RStudio, Jupyter Notebook, Advance Excel, Docker, Git, MS Office, Jira, Agile, Selenium

PROFESSIONAL EXPERIENCE

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

- Accomplished **95**% reduction in manual efforts, time, and cost by developing tool for **DOCUMENT** processing using **Machine Learning** and **Google Cloud Platform** with **8** critical features to classify, extract, translate, summarize, search documents.
- Utilized Google's Vision OCR, Form Parser API, AutoML natural language for **text detection and classification**, **human-in-loop**, NLP BART for **summarization**, Firestore data warehouse for **NoSQL** data with application deployment and monitoring on **GCP**.
- Acted as POC for multiple clients for demos and benefited clients by customizing application as per client requirements.

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

- Developed **Course Recommendation System** using Python suggesting **top 5** courses similar to employee's previous choice and collaborated with engineering and design team to incorporate this new **feature** saving 50% of employee's time.
- Predicted rating for new and unrated courses using **Linear Regression** with feature reduction (**PCA)** to achieve better results.
- Led a team of **five** and achieved **50%** reduction in testing time by developing **Selenium automation testing** framework in JAVA for Mobile App and Website of HDFC Bank with automation testing on **500** scripts daily.

ACADEMIC PROJECTS

Walmart Sales Forecasting (Python, Time-Series Forecasting)

- Estimating **28 days ahead forecasts** of products sold by Walmart in US with sales analyses by states, stores, category, dept.
- Utilizing time-series models **ARIMA**, **SARIMA** and ML models **LightGBM, LSTM, Neural Prophet** to find accurate forecasts.

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 knowledge 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 contributing feature to fraud using exploratory data analysis, performed feature engineering, predicted fraudulent transactions on highly imbalanced dataset achieving 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)

- Performed data integration(ETL) from various data sources(APIs) and economized buyers by creating model that finds highest **ROI** real estate listings and best areas to buy properties in Boston using **Causal Modeling** and **Bayesian Statistics** techniques.