Gorakhnath Nigam

Data Scientist | Bank Of America | Virtusa(xLab-Research wing) | University of Miami | IIT Guwahati

Miami, USA | jighrt16@gmail.com | +13058894767 | gorakhnigam24 | GorakhNigam

Work Experience

Data Scientist
BA Continuum (Bank Of America)
Jul 2020 – Jan 2025 | Gurugram, India

- **LLM & RAG-Enhanced Regulatory Compliance Summarization Platform:** Built an AI solution to summarize complex regulations (e.g. SR 11-7 and OCC 2011-12) into actionable guidance, cutting review time by **60%** and boosting productivity by **50%**. Developed **ingestion pipelines** (BeautifulSoup, PyPDF2, spaCy) and RAG workflow using Azure Cognitive Search with a **fine-tuned GPT-3.5** model, validated via ROUGE/BLEU and SME review. Deployed on Azure Kubernetes Service with CI/CD and Vault-secured APIs.
- **Revenue Forecasting Model:** Developed and maintained **10** Pre Provision Net Revenue **(PPNR) forecasting** models using PCA, time series, and regression models, in ***R*** and improving forecast accuracy, reducing stress-testing cycle time by **30%**, and safeguarding **$50M+ annual revenue**.
- **Fraud Detection:** Developed fraud detection for **11M+ (~20 GB)** transactions/month with Isolation Forest, clustering, and rules-based logic. Ensured model governance with PSI, AUROC, FDR, PDP, Lift values using *SAS* and built a dashboards in Tableau for the KPIs tracking. Increasing efficiency by **75%**, reducing evaluation time by **80%**, and preventing **$95M+ potential losses**.
- **Classification Model:** Built XGBoost model to classify root causes behind **UMR trade-break**, allowing faster resolution of Initial Margin and Portfolio Reconciliation breaks, cutting remediation time by **40%** and saving **$500K annually**.
- **Clustering Model:** Automated clustering with K-Means, Spectral, DBSCAN, and Gaussian Mixture Model, using advanced techniques (e.g. AIC, BIC, Gap Statistics, etc) to decide on number of clusters, reducing manual data exploration by **50%** and supporting revenue strategies with advanced ML insights.
- **ETL Tool:** Built ELT pipelines in Python, PySpark, and SQL on Delta Lake; automated ingestion, transformation, and delivery for trade related data, reducing EDA turnaround by **70%**.

Data Scientist
Virtusa
Jun 2018 – Jul 2020 | Hyderabad, India

- **Synthetic Data Generation**: Built a configurable synthetic data pipeline using metadata-driven methods and **SMOTE**, ensuring resemblance to original data. Validated similarity with hypothesis testing and deployed on **AWS EC2 with integrated storage**, cutting data provisioning time by **40%** while maintaining privacy and enabling faster model development.
- **Data Anonymization**: Automated anonymization using **perturbation** and **generalization** techniques to protect Personally identifiable information (PII). Conducted anonymity test to validate compliance with GDPR/CCPA, reducing breach risks and safeguarding sensitive data across projects.
- **Customer Segmentation & Recommendation**: Developed segmentation via **Hierarchical Clustering** and recommendation using **XGBoost** & collaborative filtering. Improved targeting and product uptake, contributing.
- **Invoice Data Extraction (OCR)**: Automated invoice parsing with **Google Cloud Vision API** and regex in Python, accurately capture key purchase details, reducing manual processing time by **60%**.
- **Sentiment Analysis (Live Tweets)**: Leveraged **NLTK** and live Twitter streaming to extract and analyze real-time sentiment on banking marketing events. Delivered insights that improved engagement rates by **15%.**

Education

MS, Data Science
University of Miami
Jan 2025 – present | Miami, USA
GPA : 4.0/4.0

B.Tech, Civil Engineering
Indian Institute of Technology(IIT)
Jul 2014 – May 2018 | Guwahati, India
CPI : 7.40/10

Skills

Programing Language & Framework
Python, R, SAS, MySql, Tableau, Power BI, Dash-Plotly, Streamlit, Pyspark, Django, Delta Lake, Git, Hadoop, AWS Sagemaker, GCP, IBM Cloud

Machine Learning and Deep Learning
Regression, Clustering Models, Ensembel Models, Classification Models, PCA, LDA, QDA, t-SNE, Neural Netowrks (ANN, CNN, RNN, LSTM, GRU) , Encoders, TF-IDF, Word2Vec, BERT, GloVe, ARIMA, SARIM, GARCH, ARCH, Copula, MCMC

Large Language Models
GenAI (OpenAI GPT, LangChain, LangGraph), LLM-Powered Chatbot Development, Retrieval-Augmented Generation (RAG), LIME, SHAP, Langfuse

Certification

IBM Data Science Practitioner Certificate
IBM | June 2025

Fundamentals of Statistics
MITx,Edx | May 2024

Probability - The Science of Uncertainty and Data
MITx, Edx | Aug 2023

Machine Learning with Python-From Linear Models to Deep Learning
MITx, Edx | May 2023

Mathematics for Machine Learning: Linear Algebra
Coursera | May 2020

Neural Networks and Deep Learning
DeepLearning.AI | Apr 2020

Machine Learning, Reinforcement Learning & Text Mining with Python
Edureka | Aug 2019

Publications

Quantifying Porosity
American Geophysical Union, Fall Meeting 2019

Abstract Link :
https://ui.adsabs.harvard.edu/abs/2019AGUFMMR 21C0082P/abstract