Hrithik Sarda

+1 (978) 654-0445 | hrithiksarda4@gmail.com | hrithiksarda | Hrithik99/Portfolio

PROFESSIONAL SUMMARY

Passionate Data Scientist with experience in data analysis, fraud detection, and building ML-driven solutions. Skilled in Python, SQL, Spark, and big data technologies to process large-scale datasets, analyze trends, develop & evaluate models. Proven ability to work in cross-functional teams & drive insights using data visualizations. Strong problem-solving skills with a focus on scalability.

EDUCATION

Northeastern University, Boston, MA — *Sep 2022 - Aug 2024*
Master of Science, Data Analytics Engineering — **GPA: 4.0**

Vellore Institute of Technology, Tamil Nadu, India — *Jul 2017 - Jun 2021*
Bachelor of Technology, Electronics and Communication Engineering — **GPA: 3.7**

TECHNICAL SKILLS

- ***Programming & Scripting:*** Python, SQL, R, UNIX /LINUX, Shell Scripting
- ***Machine Learning:*** Supervised & Unsupervised Learning, Feature Engineering, Model Evaluation metrics, Statistical Analysis
- ***ML Libraries:*** NLTK, TensorFlow, Pandas, Scikit-Learn, NumPy, TFDV, PyTorch, Airflow, ML Flow
- ***Data Analytics & Visualization:*** Data Mining, Trend Analysis, Dashboarding, Google BigQuery, Tableau, Matplotlib, Seaborn
- ***MLOps & Model Deployment:*** Apache Airflow, Kubernetes, CI/CD Pipelines, Model Risk Management, GitHub, Jira
- ***Big Data & Distributed Computing:*** Spark, Hadoop, Hive, PySpark
- ***Project Management:*** Agile (Scrum), Stakeholder Communication, Cross-functional Collaboration

WORK EXPERIENCE

Data Scientist | Northern Trust (Contractor) — Oct 2024 – Feb 2025

- **Built an AI-powered NLP system for financial document summarization**, enhancing risk assessment and fraud detection
- Fine-tuned GPT-3.5 model on financial transactions, improving contextual coherence and extractive summarization efficiency

Data Science Engineer | Bright Horizons Family Solutions — Jun 2023 – Jul 2024

- Reduced processing time by **70% for 18.3M+** contacts by optimizing geo-location matching using Haversine distance and Bing Maps API. Automated this pipeline with Apache Airflow, integrating seamlessly with Salesforce
- Developed an **XGBoost**-based anomaly detection system - This improved data flow tracking and reduced audit errors by **90%**
- Improved audience segmentation data quality by **25%** using ensemble learning models, **orchestrating incremental data loads in Snowflake from SQL Server** using Stored Procedures and Informatica PowerCenter
- Integrated SQL Server, Salesforce, and Snowflake using Informatica PowerCenter, reducing pipeline downtime

Business Intelligence Analyst | Tata Consultancy Services — Jun 2021 - Aug 2022

- Built a real-time data pipeline using **Azure Data Factory and Databricks** for efficient data ingestion and processing
- Utilized **PySpark**, reducing processing times by **60%** for terabytes of e-commerce data, enabling near real-time analytics
- Implemented a **collaborative filtering-based recommendation system** to deliver personalized insights and suggestions
- Developed a **K-Means clustering** model for e-commerce data segmentation, enabling real-time analytics and deriving insights
- Managed ETL operations with Informatica PowerCenter and Oracle/Snowflake, integrating a churn prediction model using Logistic Regression that **reduced churn by 15%**

PROJECTS

Early Sepsis Prediction ML Pipeline — May 2024 – Aug 2024

- Built an **end-to-end MLOPS pipeli**ne for sepsis prediction using Google Big Query, GCP, Apache Airflow, and Kubernetes. Automated data ingestion, validation, and preprocessing within clinical workflows
- Used **Python's MLflow library** to train and evaluate models - (Random Forest, Decision Tree, XGBoost, Logistic Regression), optimizing Random Forest on GCP's Vertex AI with **93.05%** accuracy and **94.81%** F1 Score
- Designed a CI/CD pipeline using Google Cloud Composer for continuous monitoring, automated retraining, thereby ensuring operational reliability

Melanoma Detection: Deep Learning for Accurate Diagnosis and Doctor's Insights — Jan 2024 – May2024

- Created a **deep learning framework** for melanoma detection using 33,126 dermoscopic images, incorporating preprocessing, data augmentation, and analysis to improve diagnostic accuracy
- Utilized **GANs to generate minority class images**, alongside traditional augmentation techniques, balancing the ratio to 1:2
- Achieved highest accuracy **(98.8%),** precision, and recall of **0.95** with Inception[transfer learning] model among CNN & DCNN

Exploring the Data Job Market: Analysis & Prediction — Jan 2023 – May 2023

- Analyzed job market using **Pythons matplotlib** library and applied **tokenization using word2vec** to extract key skills from JD's
- Applied clustering techniques and used elbow method to identify popular industries, regions, and job titles in the market
- Built an **Ensemble Learning Regression** model, optimizing feature selection, achieving **R^2 = 0.8456** for salary predictions