REEMA YADAV

Seattle, WA| (857)269-9106 | reemayadav1702@gmail.com | LinkedIn| GitHub

EDUCATION

Master of Science in Data Analytics Engineering, Northeastern University, Boston **Dec 2021**

Coursework: Database Management and Design, Data Structures and Algorithms, Data Warehousing and Business Intelligence, Data Visualization, Probability and Statistics, Data Mining, Machine Learning, Operation Research

TECHNICAL SKILLS

Tools: Databricks, Jupyter, Querybook, Tableau, Git, Qubole , Google Analytics, Scrum, S3, SSIS, Linux, RStudio, JIRA

Database & Languages: SQL, Python, R, Shell scripting, RDBMS (Oracle, MySQL, MS SQL), C++

ML, AI & Big Data: Data science pipeline (cleaning, wrangling, visualization, modeling, interpretation), Statistics, Regression models, LLM, RAG, Neural Networks, Hadoop, Sqoop, Hive, HDFS, Spark, ETL, Hypothesis testing, ARIMA model, A/B testing

Python: Matplotlib, NumPy, Pandas, Scikit-Learn, TensorFlow, Keras, Bokeh, Statsmodel, nltk, PyTorch, OpenAI, Huggingface

PROFESSIONAL EXPERIENCE

Expedia Group, Seattle | Data Scientist II **Feb 2022 – present**

• ***Feedback AI Engine (FAE)*** – Working on an AI-driven process with a Large Language Model (LLM) to condense customer feedback, projecting a **$25.6M** profit boost for the Product Organization. Initiated RAG research to enhance AI's data contextualization capabilities

• ***Traveler & Partner Single Data Architecture (SDA)*** – Pioneered a scalable, rolling one day SDA pipeline with **2** years of data using Querybook and Databricks Python to integrate varied structured tables for UAT and production. Engaged Databricks scheduling parallelism and data validation to ensure data quality. Delivered enriched table and Tableau facilitating data-driven decision-making for stakeholders

• ***NLP Capabilities Development*** – Engineered topic modeling, text summarization (tokenizing, lemmatizing, removing stop words), and multilingual processing, cutting analysis time from **10-12 hours to a minute** (for 1000 rows). Applied Zero-Shot learning for segregation of accessible and non-accessible feedback, boosting data categorization efficiency

• ***Reservation modification and Trust Analysis*** – Conducted a deep dive into Partner feedback to pinpoint challenges related to booking alterations and trust concerns with Expedia. Leveraged indexation, Radar Chart, and text summarization techniques to provide actionable insights to senior leadership via presentation

• ***DUET Qualtrics to S3 pipeline*** – Established Databricks data pipeline, facilitating seamless transfer of Traveler DUET data from Qualtrics to S3 through comprehensive feature engineering and transformation

• ***DUET Metric, Indexation, Feedback+, Air NDC Dashboard*** – Wireframed and created **4** Tableau dashboards tailored to meet diverse business requirements and cater to various stakeholders

SMS Group Inc, Pittsburgh | Data Scientist **Jan 2021 – Aug 2021**

• ***SSAB Asset Health*** – Utilized python to implement supervised ML technique - ARIMA for time series signal anomaly detection on **4-year** unstructured data to predict failure for furnace and burner block – decreased downtime by **40%** of furnace. Increased number of heats from **600 to 1000**

• ***Nucor Steel Force Anomaly*** – Formulated Bokeh application for time series to predict force anomaly in plant. Added numerous functionalities that cut down client's analysis time by **78%**. Delivered findings and strategic value proposition to global team head

AISkunkworks, Northeastern University, Boston | Research Assistant **Jul 2020 – Jan 2021**

• Worked on design of interpretable machine learning models - Shapely and Lime

• Conducted hands-on-experience workshops teaching EDA and ML techniques for **80+** students with Python

TATA Consultancy Services, Mumbai | Data Engineer **Dec 2016 – Dec 2019**

• ***Modeled PoC for BIRD Hadoop Project*** – Migrated historical customer transactional data of **10 years (700 TB)** in Hadoop and compressed data by **79%** deploying Snappy compression and Hive configuration

• Developed design approaches to validate findings. Automated daily data-ETL process diminishing downtime by **82%**

• Deployed MapR Hadoop clusters ensuring high availability on Production and Disaster Recovery (DR) environment

• Administered OS, Database and MapR Application upgradation on **27** Linux servers

• Devised and built Big Data Intelligence Reporting Department portal in Java for Bilingual language to extract data from Hadoop clusters for inquiry simplifying process from **1-2 days to 5-10 minutes**

• Built Grafana dashboards by ensuring data integrity, and presented insights to client improved customer satisfaction by **30%**

• Accountable for design, development, testing, documentation, and analysis of features for Hadoop project

ACADEMIC PROJECTS

Failure Detection - Predictive Analysis **Fall 2021**

Performed feature engineering on unbalanced dataset. Evaluated KNN, Logistic Regression, Gradient Boost, Random Forest, XG-Boost for precision, accuracy, recall, geometric mean, ROC curve, confusion matrix with **high** accuracy

LEADERSHIP/ACHIEVEMENTS

Executive Leadership Principles Professional Certificate by MIT

Graduate Student Assistant – Sensor Analytics | Northeastern University

Vice President of Academic Affairs - GSG| Northeastern University