REEMA YADAV

Boston, MA| (857)269-9106 | yadav.ree@northeastern.edu | LinkedIn| GitHub

EDUCATION

Northeastern University, Boston, MA **Dec 2021**
M.S. in Data Analytics Engineering
Relevant Courses: Database Design, Probability & Statistics, Algorithms, Data Mining, Data Visualization, Machine Learning

University of Mumbai, India **Jun 2016**
B.E. in Electronics

PROFESSIONAL EXPERIENCE

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

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

• KAPH Framework – Detected anomalies in Steel plant by fabricating end to end framework for analysis and prediction of time series signal using Python libraries

• Nucor Steel Force Anomaly – Designed Bokeh application for time series to predict the force anomaly in the plant from parquet file. Added numerous functionalities to the application that reduced the client's analysis time by **78%**. Presented results to team's global head and prepared a summary detailing value proposition and strategy

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

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

• Conducted hands-on-experience workshops on exploratory data analysis and ML techniques for **80+** students using Python

TATA Consultancy Services, Mumbai, India | 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%** using Snappy compression and Hive configuration

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

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

• Deployed MapR Hadoop clusters and did related developments on Production and Disaster Recovery (DR) environment. Administered OS, Database and MapR Application upgradation on **27** Linux servers

• Designed and developed the Big Data Intelligence Reporting Department portal in Java for Bilingual language to extract data from Hadoop clusters for inquiry reducing time for process from **1-2 days to 5-10 minutes**

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

TECHNICAL SKILLS

Tools: Visual Studio, Jupyter, RStudio, Tableau, Git, Azure DevOps, S3, SSIS, Pentaho, Linux

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

Machine Learning & Big Data: Hypothesis testing, ARIMA model, A/B testing, Hadoop, Sqoop, Hive, HDFS, Spark, ETL, Data science pipeline (cleaning, wrangling, visualization, modeling, interpretation), Statistics, Regression models

Python/R: Matplotlib, NumPy, Pandas, Scikit-Learn, Bokeh, Statsmodel, dplyr, tidyverse, ggplot2, tidyr, R Shiny

ACADEMIC PROJECTS

Education Quality Prediction Model **Summer 2021**

Imputed missing values with KNN in the structured data and linear interpolation to get average change to impute values for each year 2000-2018. Created a vector autoregression VAR, "minimized" AIC and BIC model on **20** selected features. Achieved **82.25%** accuracy for 2019 forecasting of Columbian education

Analysis of Sales via Warehouse Management **Spring 2021**

Performed transformation via slowly changing dimension, merge join, sort, derived columns, conditional split, data conversion to get dimension and fact table by SSIS. Data loaded into a warehouse through ETL. Designed OLAP cube

Predicting Buying Intention of Bank Customer **Fall 2020**

Performed EDA on Bank Marketing dataset using R and Tableau. Created classification model to determine whether the customer will buy a term deposit plan using algorithm- KNN, Decision Tree, SVM, Neural networks. Evaluated performance of models - confusion matrix, ROC curve and achieved **89.99%** accuracy on the logistic regression model

Supplier and Inventory Management Database Design **Spring 2020**

Cultivated an RDBMS for gaming e-commerce website with MySQL server. Designed database to Boyce- Codd normal form to curb insert, update, delete anomalies and avoid data redundancy. Implemented new order auditing via triggers, stored procedure, and programming views to track inventory

LEADERSHIP/ACHIEVEMENTS

Data for Good Hackathon Winner 2021 | JPMC

Vice President of Academic Affairs - GSG| Northeastern University