Krishna Mallik Nanduri

Bridgewater, NJ | (617)-595-8107 | nanduri.k@northeastern.edu | https://www.linkedin.com/in/nanduri-k

EDUCATION

Northeastern University, Boston, MA **Sep 2021 – Aug 2023**
Master of Science in Data Science, CGPA: 3.8/4.0

Gandhi Institute of Technology and Management (GITAM) University, India **Jun 2017 – Jun 2021**
Bachelor of Technology in Computer Science and Engineering, CGPA: 3.75/4.0

SKILLS

- **Programming**: Python(Scikit-Learn, Numpy, Scipy, Pandas, Matplotlib, Seaborn, Keras, Tensorflow), SQL, Pyspark, R, C, C++, Java
- **Statistics/Machine Learning**: Regression, Classification, Time Series, Neural Networks, Clustering, Dimensionality Reduction
- **Software/Tools**: Tableau, PowerBI, Microsoft Office, Docker, Git, Unix, JMP, SAS, AWS (EC2, RDS, SageMaker, Athena, Glue)
- **Database Systems**: AWS RDS, MySQL, IBM DB2, OracleDB, MongoDB

PROFESSIONAL EXPERIENCE

Via Separations, Watertown, MA USA **May 2022 – Dec 2022**
Machine Learning Engineer Co-op | *"Neural Networks, Predictive Models"*

- Spearheaded data-centric transition, scaling up industrial production with advanced **Time Series Forecasting** (ARIMA, SARIMA). Enhanced membrane **lifetime predictions** by integrating experimental data and **neural network algorithms**.
- Engineered a robust **Machine Learning pipeline** using **AWS SageMaker**, combining TCN (Temporal Convolutional Network) and LSTM models using Python, Keras, and TensorFlow, resulting in a remarkable RMSE score improvement from 1.9 to 0.82, optimizing membrane flux predictions
- Successfully employed **A/B testing & design** to assess the impact of varying membrane production parameters, systematically adjusting chemical configurations and operating conditions, and utilized **statistical analysis** to inform data-driven process enhancements
- Utilized SAS JMP software to apply advanced statistical methods in formulating an optimized data normalization process across membrane experiments, ensuring precise comparisons through data analysis, visualization, and statistical modeling techniques
- Collaborated with cross-functional teams to manage and optimize the **AWS RDS Relational Database** hosted on **AWS EC2 instance** using **MySQL**, implementing stored procedures and views to automate experiment logging and streamline data retrieval processes
- Advanced web development using Retool powered by JavaScript and MySQL querying, seamlessly integrating with the AWS database, facilitating efficient data management for scientists
- Streamlined data ingestion with Gmail API Cloud and Bash Script automation, creating Python scripts for efficient **data acquisition** and minimizing manual tasks

Bluebonnet Data, Minneapolis, MN, USA **Jan 2022 – Aug 2022**
Data Analyst Intern | *"Complex Visualizations, Clustering, Geospatial Analysis"*

- Directed comprehensive **data collection and cleaning** across 159 wards and precincts, ensuring impeccable **data integrity** for subsequent analyses. Utilized **Git version control** to ensure collaborative code management and seamless team collaboration.
- Executed sophisticated **geospatial data analysis** with Geopandas and Plotly, creating precision visual maps that color-coded precinct responses, offering a high-resolution insight into voter patterns
- Applied robust statistical techniques and Python programming, leveraging Pandas and NumPy, to reveal **intricate patterns and correlations** within voter demographics, contact methods, and activist affiliations
- Advanced **K-Means clustering** enabled the identification of 5 distinct voter behavior clusters, unveiling significant connections between ethnic groups and activist affiliations
- Created interactive and visually compelling **dashboards** in Tableau and Power BI to present the results of the voter behavior **analysis**. These dashboards were used by **stakeholders** to make **data-driven decisions** and **drive voter engagement strategies**
- Exhibited **agility** and **adaptability** by mastering intricate voter datasets and geospatial techniques, which culminated in **actionable insights** that influenced the company's strategic voter engagement blueprint

Verzeo (Microsoft Authorized Education Partner), Hyderabad, India **May 2021 – Dec 2021**
Data Science Intern | *"PySpark, Parallel Processing"*

- Integrated **PySpark** with MapReduce and Hadoop, revolutionizing big data processing for efficiency and scalability in **microbusiness forecasting**.
- Applied advanced ML and Deep Learning techniques, like Random Forest Regressor and LSTM, using PySpark MLlib, driving strategic decisions.
- Utilized PySpark's distributed computing for **data preprocessing**, achieving an 80% reduction compared to Python preprocessing methods
- Prioritized data quality for growth forecasting through comprehensive cleaning, exploratory data analysis, and preprocessing using PySpark's parallel processing for real-time requirements.
- Employed **feature engineering** methods (lag features, PCA) and **hyperparameter tuning**; monitored model accuracy with metrics like SMAPE, MSE, and MAE. Achieved an impressive 0.684 SMAPE score with predictive model like LSTM.
- Streamlined data processing of complex microbusiness datasets with PySpark, ensuring adaptability, precision, meaningful growth insights.
- Demonstrated a diverse skillset, from data visualization (Matplotlib, Seaborn) and **ensemble learning** to seamless PySpark-Hadoop integrations for real time processing and storage.

ACADEMIC PROJECTS

Energy Consumption Prediction | *"Boosting Techniques, Feature Engineering"* **Sep 2021 – Dec 2021**

- Developed scalable system for HVAC energy consumption prediction using a 40M-record weather and building dataset. Employed advanced feature engineering techniques to analyze weather and consumption patterns. Achieved exceptional 1.27 RMSLE score using XGBoost, LGBM, CATBoost and Random Forest models with strategic hyperparameter tuning like Bayesian Optimization and Halving Grid Search

Diabetic Retinopathy Identification | *"Computer Vision, CNN using Keras, Image Processing"* **Dec 2020 – May 2021**

- Developed a robust Convolutional Neural Network (CNN) using OpenCV's image processing and Gaussian smoothing for accurately assessing diabetic retinopathy severity from 5,000 Eye Fundus images. Achieved an impressive 80% accuracy by optimizing feature extraction through resizing to 256x256 pixels

Mental Health Analysis on Social Media Posts | *"Natural Language Processing"* **Jan 2022 – May 2022**

- Utilized Natural language processing techniques (lemmatization and POS tagging), deep learning models like Perceptron and LSTM, dimensionality reduction with Principal Component Analysis (PCA), and topic modeling using LDA to cluster Reddit posts and predict the severity of mental health situations. Achieved 80% accuracy while assessing the risk of suicide
- Employed Natural Language Processing (NLP) techniques, including TF-IDF, GloVe, and Word2Vec, for text analysis and feature extraction in Reddit posts related to mental health