# Anurag Sanjay Ghosh

Boston, MA | (857)-540-6982 | Email | LinkedIn | GitHub

## EDUCATION

**Northeastern University, Khoury College of Computer Sciences**, Boston, MA **Dec 2024**
Master of Science, Artificial Intelligence **4.00 GPA**
Relevant Coursework: Large Language Models, Reinforcement learning, NLP, Algorithms, Machine Learning.
**Teaching Assistant:** Foundation of Artificial Intelligence (CS5100)

**Mumbai University,** Mumbai, India **June 2019**
Bachelor of Engineering, Computer Engineering

## TECHNICAL SKILLS

**Programming/Scripting:** Python, JavaScript, Java, TypeScript, C++, C.
**Databases and Cloud Service:** Google Cloud Platform, Big Query, AWS, Sagemaker. Azure, MySQL, Microsoft SQL Server
**Tools & Frameworks:** Tableau, PowerBI, NodeJs, Python (Keras, PyTorch, OpenCV, TensorFlow, Pandas, Scikit-learn, matplotlib, Gensim, NLTK, Ludwig, Vertex AI, spacy, pandas, scipy, Huggingface), R, Docker, GitHub, Streamlit, PySpark, Superset.

## WORK EXPERIENCE

**Ribbon Communications,** Westford, MA **May 2024 – August 2024**
***Data Science Intern***

- Developed a dockerized **URL categorization** service utilizing **Selenium** web scraping engine. Leveraged finetuned **RoBERTa** to perform categorization on scraped webpage & meta data with a **Zero Shot DeBERTa** model for further sub categorization.
- Achieved a F1 score of **85%** & reduced unknown/new URLs categorization & processing time on client (AT&T) side by **70%**.
- Designed an algorithm to preprocess, parse (SPELL) and cluster unstructured logs based on semantic similarity and temporal occurrence patterns. Utilized **isolation forest** on occurrence frequencies to identify anomaly windows in the log's timeline
- Integrated preprocessing pipeline with a **Streamlit** app, allowing users to zoom into anomaly windows & analyze log clusters to identify potential trigger events. Enabled **chat with logs** of interest using an **OLLAMA** chatbot for further insights.

**Universal Music Group,** New York City, NY **July 2023 – December 2023**
***Data Science Analyst Intern***

- Designed **LightGBM** based **Customer LTV Models** on **AWS Sagemaker** for predicting purchase propensity & identifying superfans (high value customers). Processed **millions of transactions** to train model, & interpreted predictions using **Shapley**.
- Elevated leadership's awareness of **data-driven** approach to fanbase analysis & customer valuation leading to its prioritization.
- Prepared a **feature extraction module** for raw emails of artist's release campaigns. Devised a **BERT** based **active learning technique** to achieve agile labeling & training for models to tag emails with predefined categories & unsupervised labels.
- Achieved a **10% increase** in performance of downstream **in-production,** engagement assessment models with the features.
- Developed a **Bayesian Linear Regression** model to predict artist's revenue from factors like music, merch release & streaming data. Provided business with a **Streamlit** portal to tweak various factors & use the model to analyze their effect on the revenue.

**WINES Lab Northeastern University,** Boston, MA **January 2023 – September 2023**
***Machine Learning Research Assistant***

- Explored **machine learning based radio fingerprinting** by training neural networks like **Alex net** & **Vision Transformers** on sequences of in-phase (I) and quadrature (Q) data from **LoRa** devices to identify the devices in varying external scenarios.
- Achieved **99% accuracy** across all device configurations when training and testing data was collected on the same day.

**TIAA GBS,** Mumbai, Maharashtra **July 2019 - July 2022**
***Analyst, Software Developer***

- Lead a team of 4 developers overseeing **development & production** of **3** spring boot microservice applications.
- **Independently** supported & collaborated with business on production deployment for **retirement income evaluation apps**.
- Achieved a **100%** success rate with more than **20** releases within a year by making applications **rapid release** compatible with **OpenShift** integration. Reduced production deployment time from several days to few hours.
- Analyzed data and generated reports for insurance agents by creating KPI dashboards using **Power BI**. Streamlined user session and engagement tracking through automated **SQL Batch Procedures**, resulting in reduced efforts for the insurance team.

## ACADEMIC PROJECTS

**Store Sales Prediction on Time series Data** **February 2023 - April 2023**

- Programmed **LSTM** and **XGBoost** regression models to predict sales of a store from previous time series data of sales.
- Performed **EDA** to find the periodicity & correlation of sales with factors like oil prices, holiday events and promotions.
- Achieved a R-squared score of **0.98** and **0.82** on testing data, with XGBoost and LSTM models respectively

## PUBLICATIONS

"Hybrid Image Encryption Technique Using Genetic Algorithm and Lorenz Chaotic System" ITM Web of Conferences 32, 03009 (2020), https://www.itm-conferences.org/articles/itmconf/pdf/2020/02/itmconf_icacc2020_03009.pdf