SETHU S

Data Scientist | Statistician | AI/ML Engineer

sethus4791@gmail.com | **github.com/itsmesethus** | **Portfolio: itsmesethus.github.io/github-porfolio/**
Phone: +91-8122815260 | Bangalore, Karnataka | linkedin.com/in/sethus4791

SUMMARY

AI/ML Engineer with 3 years of experience building production-grade ML and statistical systems at Nielsen. Delivered audience weighting engines, anomaly detection pipelines, and TV viewership estimators that directly power US national TV ratings using Python, PySpark, SQL, and Databricks. Supported MRC audits by Ernst & Young with zero findings across 4 validation pipelines.

EXPERIENCE

AI/ML Engineer | **Nielsen, Bengaluru, India** ***Jul 2024 – Present***

- Cut reporting friction for upper management by turning 10 months of Big Data vs Panel weight data into a single, decision-ready dashboard, reducing time-to-insight across US demographic segments.
- Maintained 100% client query resolution rate across 8+ critical IPF weight anomaly escalations by diagnosing hidden root causes in low-sample DMAs that standard checks missed.
- Prevented a production-level rating failure by catching critical Universal Estimates discrepancies during Time Period Rating Engine UAT before deployment, safeguarding Nielsen's national TV measurement accuracy.
- Directly satisfied complex client compliance requirements by owning Custom ACR (Vizio-Roku) weighting support and technical audits end-to-end.

Statistician | **Evoscien, Bengaluru, India** ***May 2023 – Jun 2024***

- Improved decision accuracy by 25% by building a reliable statistical foundation from fragmented insect population data, enabling consistent, evidence-backed research outcomes.
- Partnered with Entomology and Engineering teams to deliver data-driven solutions that improved operational efficiency by 15% across field and lab workflows.
- Surfaced hidden patterns in insect behavior and population trends using t-tests, ANOVA, chi-square, and non-parametric tests at 95% confidence intervals.
- Produced 20+ interactive dashboards and reports that translated complex statistical findings into clear, actionable insights for both technical and non-technical audiences.

PROJECTS

a) Audience Weighting Feature Selection Engine

- Reduced 50+ candidate variables down to the most predictive weighting characteristics for Nielsen's national audience estimation pipeline, using weighted stepwise regression with adjusted R^2 to maximize model fit.
- Ensured 8+ minority demographic subcategories (Black, Asian, Hispanic) were accurately represented by applying ANOVA-based feature ranking, correctly prioritizing statistically significant variables despite data sparsity and imbalance.
- Delivered finalized weighting characteristics and implementation guidelines to 6+ downstream teams, ensuring methodological compliance across the 2025-2026 US National TV rating cycle.

b) Big Data Quality & Exclusion Framework

- Built an Isolation Forest-based anomaly detection pipeline to identify and exclude corrupted tuning data caused by natural disasters (tornadoes, hurricanes) across 210 DMAs, preserving viewership signal quality for national TV measurement.
- Engineered dynamic 3-sigma exclusion thresholds across 5 Big Data providers using clean-day baselines, establishing reliable benchmarks for Installed, Intab, and Household Tuning Minutes metrics.
- Delivered validated exclusion thresholds to the engineering team for production integration as data quality gates in Nielsen's Samples selection pipeline for the 2025-2026 rating cycle.

c) Big Data Viewership Rating Estimator

- Reduced TV ratings reporting latency from 6 days to 1 day by developing a Big Data-driven estimation system that predicts panel-based viewership ratings, enabling significantly faster audience insights for US TV markets.
- Engineered predictive features from historical viewership data including lag-based temporal signals, station-level metadata encodings, and market-level aggregations to capture complex viewing patterns across diverse US TV stations.

- Delivered stable, reliable predictions across 80+ major US TV stations by training and hyperparameter-tuning a LightGBM regression model, evaluated using RMSE and cross-validated across multiple market segments.

d) Big Data Media Landscape — Presence of Attributes

- Future-proofed Nielsen's Big Data weighting pipeline by replacing legacy media attributes (Cable Plus, vMVPD, OTA) with modern presence-based features (Multichannel, Antenna, Internet, SVOD), aligning national TV measurement with the industry's shift toward streaming.
- Quantified the weighting impact of the framework migration by rigorously comparing IPF-generated weights under old vs new attribute schemas across both NPM and Big Data+Panel pipelines.
- Presented methodology enhancements to the MRC audit team (Ernst & Young) in a formal walkthrough, achieving zero audit findings across the updated weighting pipeline.

EDUCATION

M.Sc. in Statistics | Bharathiar University, Coimbatore **| 8.61/10** ***Sep 2021 – May 2023***

B.Sc. in Statistics | Arignar Anna Govt Arts College, Villupuram **| 9.39/10** ***Jun 2018 – May 2021***

Relevant Coursework: Descriptive Statistics, Sampling Theory, Probability Theory, Statistical Estimation Theory, Statistical Quality Control, Multivariate Statistical Analysis, Distribution Theory, Econometrics, Statistical Inference, Programming in R, Design of Experiments, Stochastic Processes.

SKILLS

Programming: Python, SQL, R, PySpark

Tools: Microsoft Power BI, IBM SPSS, MINITAB, STATISTICA, Microsoft Excel, MySQL, Jupyter Notebook, Git, GitHub, Azure Databricks, AWS S3

Libraries: Pandas, NumPy, Matplotlib, Seaborn, Plotly, SciPy, Scikit-Learn, TensorFlow, Keras, Statsmodels, Pingouin, Streamlit

Core Competencies: Data Cleaning, EDA, Feature Engineering, Feature Selection, Data Visualization, Outlier Detection, Correlation Analysis, A/B Testing, Ad Hoc Analysis, Model Evaluation, Model Deployment, Hyperparameter Tuning

ML & Statistics Expertise: Regression, Classification, Clustering, Predictive Modeling, Quantitative Analysis, Statistical Modeling, XGBoost, LightGBM, Ensemble Methods, Deep Learning (ANN, CNN, RNN, LSTM, GRU), NLP, Transformers, Time Series Forecasting, LLMs & RAG(Basics).

CERTIFICATIONS

- IBM Data Science Professional Certificate — IBM / Coursera
- Machine Learning Specialization (Supervised ML, Advanced Learning Algorithms, Unsupervised ML & Recommender Systems) — DeepLearning.AI / Coursera
- Ensemble Methods in Python — DataCamp
- Sequences, Time Series and Prediction — DeepLearning.AI / Coursera
- Feature Engineering for Machine Learning in Python — DataCamp
- Data Analytics with Python — NPTEL / Swayam
- Databases and SQL for Data Science with Python — IBM / Coursera