

Komal Sairam Reddy Bhimireddy

San Francisco, CA | +1 (669) 204-6802 | komalbhimireddy@gmail.com | [Portfolio](#) | [LinkedIn](#) | [GitHub](#)

SUMMARY

Machine Learning Engineer with experience building data-centric ML pipelines and fine-tuning ASR models, achieving measurable gains in production systems. Strong background in statistical modeling and large-scale data analysis, with peer-reviewed research experience. Proven ability to translate data quality improvements and rigorous evaluation into reliable, real-world model performance.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer Intern – Suki AI, Redwood City, CA

Jul 2025 – Oct 2025

- Owned an end-to-end **ASR data optimization pipeline** spanning EDA, waveform-level deduplication, targeted augmentation, and Whisper fine-tuning, improving the reliability of clinician-facing voice commands and establishing a reusable baseline workflow.
- Discovered and removed **1,087 hidden duplicate audio samples** via **waveform SHA-256 hashing**, eliminating training data leakage and increasing effective dataset diversity without additional data collection cost.
- Analyzed dataset bias and intent skew, then performed **targeted augmentation** (+1,700 synthetic commands) using **TTS to rebalance speaker gender and underrepresented command types**, improving coverage of critical but failure-prone workflows.
- Fine-tuned Whisper-Medium on the optimized dataset, achieving an **8.5% relative WER reduction** and **6.7% token-level recall improvement**, directly reducing retries and improving first-try recognition accuracy in real clinical usage scenarios.
- Evaluated multiple fine-tuning strategies and identified **catastrophic forgetting** in sequential training, leading to the selection of a **combined-training baseline** and informing future ASR training guidelines around data mixing and regularization.
- Tech:** Python, Whisper, Hugging Face, ASR evaluation (WER/CER), TTS, GCP, Git, data pipelines.

PUBLICATIONS

Forecasting Gold Returns Volatility Over 1258–2023: The Role of Moments | [Link](#)

September 2025

Applied Stochastic Models in Business and Industry, Wiley, 2025

- Developed Bayesian time-varying quantile regression models to derive tail-risk, skewness, and kurtosis measures from 766 years of gold return data, demonstrating significant improvements in out-of-sample volatility forecasting over autoregressive benchmarks.

RESEARCH EXPERIENCE

Research Assistant – BRFSS SMART Analysis: Binge Drinking and Frequent Mental Distress

Feb 2025 – Present

Advisors: Dr. Anandamayee Majumdar (Mathematics, SFSU), Dr. Muntasir Masum (Epidemiology & Biostatistics, UAlbany)

- Constructed an analytic dataset of **726,000+ BRFSS respondents** by harmonizing variable definitions across years and preserving survey weights, strata, and PSUs to ensure valid inference under the complex sampling design.
- Implemented a **stepwise survey-weighted logistic regression** framework to evaluate the association between binge drinking and frequent mental distress, sequentially introducing demographic, socioeconomic, behavioral, and health-related covariates.
- Identified socioeconomic status and smoking as key confounders, with the fully adjusted model showing a small but stable positive association (**OR = 1.04**); model fit improved substantially (AIC decreased by **100,000**, **pseudo-R²** increased to **0.110**).

PROJECTS

Large-Scale Sentiment Analysis (Python, scikit-learn, NLP, TF-IDF, SVM, Random Forest, Gradient Boosting, Imbalanced Data Handling)

- Built a multi-class sentiment classification pipeline on **205k+ e-commerce reviews**, performing text normalization and **TF-IDF vectorization**, and benchmarking **SVM, Random Forest, and Gradient Boosting** models using F1 and Recall.
- Designed controlled experiments to evaluate **class imbalance handling strategies** (Random Under-Sampling, Tomek Links, SMOTE), quantifying their impact on decision boundaries and generalization rather than assuming preprocessing improvements.
- Demonstrated that sampling degraded performance in this regime, with **SVM on raw data outperforming all resampled variants (F1 = 0.936, Recall = 0.942)**, highlighting the importance of dataset-aware modeling decisions over heuristic balancing.

Music Recommendation System (Python, scikit-learn, Spotify API, Regression, Classification, Cosine Similarity, Feature Engineering)

- Built an end-to-end music recommendation pipeline by ingesting **5.3k+ tracks from Spotify playlists**, extracting audio features and metadata, and performing feature scaling and preprocessing for downstream ML models.
- Formulated song popularity prediction as both a **regression and a binary classification problem** to compare continuous scoring vs decision-based modeling, selecting models based on task-aligned metrics (Regression R² = **0.269**, Classification Recall = **0.762**).
- Developed a **content-based recommendation engine** using **cosine similarity** over audio features, leveraging predicted popularity signals to rank and optionally filter personalized song recommendations.

EDUCATION

San Francisco State University, San Francisco, CA

May 2025

Master of Science, Statistical Data Science. (GPA: 3.69/4.0)

Coursework: Data Mining, Probability & Statistics, Advanced Probability Models, Experimental Design, Computational Statistics, Statistical & Machine Learning, Multivariate Statistical Methods

TECHNICAL SKILLS

Programming: Python (NumPy, Pandas, scikit-learn, PyTorch, Hugging Face, TensorFlow), R, SQL, Shell Scripting

Machine Learning: Regression, SVM, Random Forest, Gradient Boosting, Neural Networks, Transformers, Recommendation Systems

Statistical Modeling: Bayesian Methods, Survey-Weighted Regression, Quantile Regression, Confounding Analysis

Data Preparation: Data Cleaning, Feature Engineering, Data Augmentation, Data Harmonization, Data Validation

Model Evaluation: Cross-Validation, F1/Recall/Precision, WER/CER, Error Analysis

Cloud & Tools: Google Cloud Platform (VMs, Cloud Storage), Git