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 and model optimization pipeline** spanning dataset EDA, waveform-level deduplication, targeted augmentation, and Whisper fine-tuning, improving reliability of clinician-facing voice commands in production.
- **Built automated waveform-level deduplication into the ASR data pipeline**, detecting and removing **1,087 hidden duplicate audio samples**, preventing data leakage and improving dataset diversity **without additional data collection or labeling cost**.
- **Analyzed intent imbalance and failure patterns**, then applied targeted TTS-based augmentation (+1,700 synthetic commands) to expand coverage of underrepresented clinician workflows and **reduce errors in previously failure-prone command categories**.
- **Fine-tuned Whisper-Medium on the optimized dataset**, achieving an **8.5% relative reduction in WER** and **6.7% relative improvement in token-level recall**, leading to **fewer recognition failures and retries in real-world usage**.
- Modeled the workflow impact of ASR accuracy improvements, estimating **1,100+ clinician-hours saved per month** across active users, **improving operational efficiency at scale without changes to clinician behavior**.
- **Established a repeatable, evaluation-driven ASR improvement pipeline**, enabling faster iteration and more predictable releases by tightly coupling data quality checks, augmentation, and model evaluation into a single workflow.
- **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

- Examined the predictive role of higher-order moments (tail risks, skewness, kurtosis, leverage) for gold volatility using 766 years of historical data, providing the first long-horizon evidence that moments-based models outperform 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)

- **Examining whether binge drinking has an independent association with frequent mental distress** using pooled, nationally representative BRFSS data (2013–2019), addressing ambiguity in prior studies driven by demographic and behavioral confounding.
- **Applied a stepwise survey-weighted logistic regression framework** to isolate the effect of binge drinking under the BRFSS complex sampling design, revealing a **reversal of the crude association** after socioeconomic and behavioral adjustment.
- **Demonstrated that the observed association between binge drinking and mental distress is sensitive to confounding**, with adjusted analyses indicating a modest independent effect of binge drinking and **showing disability and smoking to be dominant**.

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 $\mathbf{R^2 = 0.269}$, 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.

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

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