PRATEET MISHRA

+1 (510) 906-5850 | prateetm@usc.edu | LinkedIn | GitHub | Portfolio | Los Angeles, CA

SUMMARY

Machine Learning Engineer focused on **production recommenders/ranking**: real-time feature pipelines/stores, **sub-second** APIs, and A/B (offline **nDCG**, online **CTR/conv**) delivering **+25% engagement**, **−20% feature latency**, **+30% throughput**.
Fine-tuned medical-QA **LLM** with **RLAIF + reward modeling**: accuracy **+4.7%**, BERTScore **+4.8%**, BLEURT **+7.8%**.

EDUCATION

University of Southern California **Jan 2024 - Present**
Master's, Computer Science, GPA: 3.4/4.0
- Relevant Coursework: Analysis of Algorithms, Web Technologies, Information Retrieval & Web Search Engines, Computer Security
- **Technical Student Lead**, Information Technology Services - mentored **50+ assistants**; primary POC with ITS engineers

Symbiosis University of Applied Sciences **Aug 2019 - Jun 2023**
*Bachelor of Technology, Computer Science and Information Technology, GPA**:** 9.23/10*
- Relevant Coursework: Object-Oriented Design, Database Management, Cloud Computing

EXPERIENCE

NICT Technologies Pvt. Ltd. **Sep 2022 - Dec 2022**
Machine Learning Engineer Intern *Indore, India*
- Coordinated with a 6-member team to build a content-based NFT recommender (profile + item similarity, cosine) in React, increasing engagement by ~25%.
- Engineered real-time feature pipelines from on-chain mint/list/bid events to refresh user/item vectors, cutting feature latency by ~20% and boosting throughput by ~30%.
- Deployed a low-latency ranking service (API inference cached per session) delivering sub-second recommendations during protected browsing and transactions, supporting 100+ trades/day.
- Ran A/B analyses on checkout guardrails using event telemetry; surfaced high-leak failure patterns and shipped fixes that cut failures by 26% and raised completion by 15%.

Ypsilon IT Solutions Pvt. Ltd. **Jun 2022 - Sep 2022**
Machine Learning Engineer Intern *Remote, India*
- Architected a real-time Scrap Auction system (Django + MySQL) with RBAC (admin/seller/user) and ops dashboards, capturing clean auction/bid telemetry for model monitoring; handled 200+ concurrent bids and cut manual ops by ~40%.
- Shipped low-latency DRF APIs and live sync for listing/bidding/checkout, maintaining online/offline feature parity and a <200 ms inference budget for ranking hooks; reduced page latency by 25% and improved bid response +30%.
- Implemented collaborative-filtering recommendations backed by a feature store (bidding history/preferences), served via API with session caching for low-latency delivery.
- A/B-validated the rollout, tracking nDCG offline and CTR/conversion online; achieved +18% repeat participation and +12% average bid.

ACADEMIC PROJECTS

SmartMedAI
- Fine-tuned an LLM for medical Q&A with RLAIF on a curated clinical Q&A set, strengthening healthcare domain recall.
- Built a **reward model** (evaluator LLM) to score candidate answers and guide RL updates for higher-quality responses.
- Optimized via RL and validated on 200 USMLE-style questions**:** DeepSeek-R1 accuracy 0.769 → 0.805 (+4.7%)**,** BERTScore F1 0.621 → 0.651 (+4.8%)**,** BLEURT 0.527 → 0.568 (+7.8%)**.**

Transfer Learning for Image Classification
- Built a **transfer-learning pipeline** in **TensorFlow/Keras** (ResNet50/101, EfficientNet-B0, VGG16) with **OpenCV** preprocessing, on-the-fly augmentation, and **hyperparameter sweeps** for robust model selection.
- Established **reproducible ML workflow**: stratified train/val/test splits, early stopping, checkpointing, and metric logging to compare architectures fairly.
- Evaluated with **precision, recall, F1, ROC-AUC**; selected **EfficientNet-B0** as best model with **≈90% top-1 accuracy**, **highest F1**, and **lowest test loss**.

Real State Price Prediction
- Devised an end-to-end **price prediction pipeline** in Python (**pandas, scikit-learn**): feature engineering (BHK, sqft, locality), outlier handling, and **k-fold validation**, achieving **R^2 = 0.87** on **10k+** Bengaluru listings.
- Shipped a Flask inference service with an interactive UI for real-time estimates, cutting manual property evaluation time by **~60%**.

TECHNICAL SKILLS

- **Languages**: Python, SQL, JavaScript/TypeScript, HTML5, CSS3
- **Frameworks & Tools**: Machine Learning Framework, TensorFlow, LLM, GenAI, NLP, React, Node.js, REST API, MongoDB, React, Node.js, Google BigQuery, Google Cloud Platform, AWS, Data Engineering, Analytical Thinking, Client-facing Skills