Puneet Ludu

Machine Learning Engineer · 13+ years in ML/AI systems
puneet.ludu@gmail.com | New York, NY | +1-(716) 867-4344 | puneet.io

https://github.com/puneetsl
https://www.linkedin.com/in/puneetsl
https://www.kaggle.com/puneetsl
Google Scholar

Experience 13+ years

Zillow (Zestimate), Machine Learning Engineer, *Sep 2021 – Present*, Remote **4Y**

Listing IQ: Interactive CMA Platform *(Django, DocumentDB, PyTorch)* [Real-time Valuations + Embeddings]

Led proof-of-concept and core architecture for platform now live as Listing IQ CMA – AI-powered tool enabling agents to customize property comparisons with map-based filtering, editable valuations, and amenity-level explanations. Designed APIs integrating real-time Zestimate valuations, property embeddings, and comparative analysis. Mentored junior engineer through production launch.

Impact: *POC → shipped product powering agent pricing decisions; foundation for new agent revenue products*

Active Listing Comps Engine *(Apache Spark, Metaflow, H3 Geospatial)* [Similarity Scoring + Batch/REST APIs]

Architected and built similar listings comparison engine from ground up – batch pipeline and real-time API using H3 geospatial indexing and customizable ranking algorithms. Processes 3M+ listings, powers Zillow Showcase dashboards (8M monthly views) helping agents demonstrate listing performance vs non-Showcase homes. Led ALC migration to DPDS, production in 5 weeks from project start. Later led algorithm improvements.

Impact: *5 weeks to production; 95% complaint reduction, coverage 90%→98.5%; metrics cited in sales, marketing, and investor communications*

Infrastructure & Engineering Leadership *(Terraform, AWS, Metaflow, Docker)*

Built Valuation API from POC to production – FastAPI service handling ∼6K requests/day with zero P2 alerts since launch. Led Zestimate 6.6 deprecation (Redis→ODP migration, legacy system shutdown), saving $350K annually. Drove Z7.1 point release to completion – created MR tracking system, led ETL optimizations. Completed CI/CD pipeline modernization across all team services with zero production incidents.

Impact: *$500K+ combined annual savings, 61% alert reduction (641→249 YoY)*

AI/ML Initiatives (In Progress) *(PyTorch, CLIP, LangChain, Databricks, MLflow)*

Leading Image Embeddings project – designing DualLossAutoEncoder to predict property condition from CLIP embeddings (3.3M listings) for Zestimate integration (DRI, cross-team coordination with computer vision and valuation teams). Built end-to-end LLM-powered customer care tool (React + LangChain), first AI assistant deployed to Zillow customer support, enabling agents to generate personalized Zestimate explanations. Led ZNN-ETL migration to Databricks – identified 8x latency and 10x cost regression in initial migration, optimized to match production baselines before deployment. Modernizing core Zestimate neural network pipeline.

Impact: *Advancing visual ML and LLM capabilities for valuation products; first LLM tool deployed to customer care*

Mentorship & Technical Leadership

Managed summer intern (2023) – designed project plan, weekly check-ins, received "strongly favorable" feedback. Mentored 4+ engineers across Neural RentZestimate, Call Insight deployment, and onboarding. Created RFC templates adopted by the team. Conduct technical interviews, lead code reviews, and coordinate cross-team design discussions.

Match Group (OkCupid), Machine Learning Engineer, *May 2020 - Sep 2021*, New York City **1.5Y**

Discount Optimization *(Python, Keras, TensorFlow, Weights and Biases)* [Wide&Deep]

Owned end-to-end ML pipeline for subscription discount optimization – feature engineering, model training, A/B testing, deployment, and production monitoring. Researched and implemented novel uncertainty modeling technique to address model instability.

Impact: *6% overall revenue increase through A/B tested pricing models*

FactSet, ML Engineer → Senior ML Engineer, *Apr 2015 - May 2020*, New York City **5Y**

ML-Powered Financial Data Extraction *(Python, TensorFlow, Keras, Sagemaker)* [CNN, ELMo, BiLSTM]

Led multiple ML initiatives: (1) Speaker identification system for earnings calls using spectrograms and CNNs, (2) Private company fact extraction from 1.6M websites using ELMo/BiLSTM. Rewrote MLangID language identification service. Led machine translation infrastructure (Polish SMT achieving BLEU 69.10).

Impact: *20% reduction in human-hours for earnings call processing, automated extraction from millions of documents* [📄]

Financial Document Search & Ranking Systems *(Apache Spark, Java, Python)* [Distributed Trie, N-gram LM, Vector Space]

Led team of 3 engineers on Document Screening – built autosuggestion and concept similarity systems. Created FingerPrinter deduplication service (10× response improvement: 1000ms→100ms). Architected Formula Lookup using distributed trie and n-gram language models on Spark.

Impact: *Improved formula ranking from 5.6 to 2.3, 66% faster document processing, powered StreetAccount trending news*

Technical Leadership

Established engineering best practices: Jenkins CI, comprehensive test suites, documentation standards. Mentored new hires and junior engineers. FingerPrinter became the model Java project within the ML group.

Tata Research Development and Design Centre, ML Research Engineer, *July 2011 - July 2013*, India **2Y**

Event Detection in Time Series *(Java, Python, RapidMiner)* [SVM - RBF]

Wrote an algorithm based on Shape Context for finding frequently occurring patterns and events, with as good results as SAX, DTW etc. with **7%** better results in the particular domain of car sensors.

Data Harmonization Framework (DHF) *(Java, Apache Pig)*

Implemented an ETL framework that exploits the power of map-reduce and big-databases to fuse incongruous enterprise data from disparate sources in near real time.

Skills

Languages	Python · Java · C/C++ · Bash · SQL
ML/AI	PyTorch · TensorFlow · CLIP · LangChain · RAG · A/B Testing · Monitoring
Data & Infra	PySpark · Databricks · MLflow · Metaflow · FastAPI · Django · Docker · Kubernetes · Terraform
Cloud & CI/CD	AWS (S3, EC2, SageMaker) · GitLab CI · W&B · Pinecone
Leadership	System Design · Technical Interviews · Mentoring · RFC Authorship · Cross-team Coordination

Publications

Google Scholar profile

Inferring Latent Attributes of an Indian Twitter user using Celebrities and Class Influencers	ACM Hypertext 2015
Inferring gender of a Twitter user using celebrities it follows	CORR 2014
Architecture for Automated Tagging and Clustering of Song Files According to Mood	IJCSI, 2010

Education

Master of Science in Computer Science, State University of New York, Buffalo, NY 2014

B. Tech. in Computer Science and Engineering, JIIT, India 2010

Open Source & Community

Organizer @ MUFin	Program committee member, paper reviewer at top ML conferences – Workshop on Modeling Uncertainty in the Financial Sector *(AAAI 2023, ECML-PKDD 2022)*
Lotion	Unofficial Notion.so Desktop app for Linux *(2K+ GitHub stars / 60K+ Clones & Downloads)*
Romadeva	Tool to convert Roman script to Indic(Devanagari) script *(Used by Translators Without Borders)*
jTextBrew	A JAVA library for fuzzy string matching, based on TextBrew algorithm by Chris Brew
Quena	Question and Answering system – Indexed 1.6 Million Wikipedia documents, designed a question parser and a ranking algorithm based on popularity. *(Apache Solr, NER, POS tagger)*