Prashant Surupsing Gavit

Phone: +1-669-294-1874 **Email**: prashat.gavit@sjsu.edu **LinkedIn**: *prashant-gavit-3b3540118* **Github**: *prashantgavit*

Summary

Data Scientist with 7+ years in U.S. healthcare and logistics, skilled in machine learning, statistics, deep learning, software engineering, and team leadership. Proven success in developing scalable AI/ML solutions, including a multi-objective recommendation system and patient risk assessment tools.

Education

- **San Jose State University** — *San Jose, USA*
  M.S. Artificial Intelligence - GPA 3.775/4 — *Jan 2024 - Dec 2025*
  Relevant Coursework: Machine Learning, AI & Data Engineering, Recommendation Systems, Data Structure & Algorithm, DBMS
- **Indian Institute of Technology Madras (IITM)** — *Chennai, India*
  BTech and MTech- GPA 3.4/4 — *August 2011 - May 2016*

Experience

- **Blackbuck** — *Bengaluru, India*
  Senior Data Scientist — *April 2022 - July 2023*
  - Developed a **multi-objective recommendation system**, leveraging a **contextual multi-armed bandits**. Enhanced recommendation diversity while maintaining NDCG metric and reduced A/B testing time by 30%.
  - Developed a real-time GPS outlier detection system using a **Dynamic Kalman filter**, achieved 95% noise detection with only 0.1% false positives. This contributed to an 18% increase in active user count.
  - Developed a semantic representation of product entities using a **GloVe** model and incorporated these semantics as recommendation model features, improving the NDCG metric of the recommendation system by 13%.
  - Developed data processing pipelines for feature engineering using **SQL, AWS Athena, S3, and Apache Airflow**, and built a deployment framework with AWS SageMaker and MLflow, reducing ML model deployment time by 80%.
- **Innova Solution** — *Chennai, India*
  Tech Lead — *Sep 2021 - March 2022*
  - Built a centralized data lake for de-identified U.S. healthcare data from multiple sources using **AWS Athena, AWS Lake Formation, and data mesh architecture**, resulting in a 31% increase in data usage.
  - Integrated the data lake with **AWS Sagemaker, Superset, Power BI, and Tableau**, enabling data science teams to build ML models and visualizations, reducing overall data discovery and model development time by 38%.
- **Innovaccer** — *Noida, India*
  Senior Data Scientist — *June 2016 - August 2021*
  - Led a team of data scientists to develop **Patient Identity** and **Risk Management** solutions, collaborated with product and customer success managers to identify customer needs, defined project initiatives and roadmaps, and delivered end-to-end ML projects, contributing nearly **$10 million Annual Recurring Revenue (ARR)**.
  - Developed an index with **Principal Component Analysis** and Google Maps data to assess social vulnerability risk, enabling better patient prioritization and increasing customer adoption of the risk management solution by 40%.
  - Trained a bidirectional **LSTM** model on historical clinical data to predict chronic disease onset, achieving an AUC-ROC of 0.85. This enabled the accountable care organization to prioritize high-risk patients and reduce the cost of care by 12%.
  - Led the development of a Healthcare Data Insights Dashboard in **Power BI**, providing insights on **patient risk, care gaps, utilization cost and readmission rates**. Empowering Accountable Care Organization leadership to reduce the cost of care by 8%.

Academic & Research Projects

- **Enhancing Recommendation Fairness Through Balanced Data Generation** — *July 2024 - current*
  - Developed an architecture integrating the **Variational Autoencoder** and **Diffusion model** to generate synthetic data for the female population, training a **Neural Collaborative Filtering** model that achieved 8.7% reductions in demographic parity compared to the baseline model. **Advisor**: Professor Magdalini Eirinaki
- **Improving Geolocation Prediction with Deep Reinforcement Learning - Git** — *June 2024 - August 2024*
  - Developed a model to optimize image selection for geolocation prediction using the Deep Deterministic Policy Gradient (DDPG) model. Implemented a CNN-based actor to select high-quality images and a CNN-based critic to evaluate their suitability. This approach improved prediction accuracy by 18% by filtering out low-quality images.
- **Classification explainability analysis on Llama models - Git** — *June 2024 - August 2024*
  - Integrated the methods from the paper "**Explaining by Removing: A Unified Framework for Model Explanation**" with the **LLaMA 3 8B** model to enhance the **explainability** of LLM models for predictions in COVID-19 fake news classification task. **Advisor**: Professor Vishnu S. Pendyala

Skills

- **Languages:** Python, C, C++, R, MATLAB, SQL, Scala
- **ML Frameworks & Tools:** PyTorch, TensorFlow, scikit-learn, NumPy
- **Big Data & Cloud:** AWS, GCP, Azure, Hadoop, Spark
- **Deployment & DevOps:** Docker, Git, Apache Airflow, CI/CD