Jasmine (Feifan) Guan

Albany, CA 94706 | (609) 907-0896 | feifanjasmineguan@berkeley.edu | LinkedIn | MyImpact

Data Scientist with 3+ years of experience delivering scalable NLP and ML models
that drive measurable business impact and cost savings.
Expert in Python, SQL, parallel processing, and cross-functional communication.
Passionate about using data to solve ambiguous problems and elevate decision-making.

EDUCATION

University of California, Berkeley **Aug 2025 - May 2026**
Master of Analytics - Industrial Engineering and Operations Research
- **Courses:** Machine Learning, Database Design, Optimization, Statistics, Risk Management & Simulation

University of California, San Diego **Sep 2018 - Mar 2022**
Bachelor of Science: Data Science & Mathematical-Economics | **GPA**: 3.8/4.0

SKILLS & CERTIFICATES

- **Languages:** Python, SQL, JAVA, Spark, Scikit-learn, MLFlow, Tensorflow, LLMs, Databricks, AWS, GCPSnowflake,
- **AI/ML**: Machine Learning, Business Intelligence, Econometrics, Statistical Analysis, Financial Data Analysis, Data Structures, Database Warehouse, Neural Networks, Production Deployment, MLOps, Project Management
- **Certificates**: AWS Certified ML Specialty, AWS Cloud Practitioner, Deep Learning.ai Certificate, IBM DS Professional

PROFESSIONAL EXPERIENCE

UC Berkeley - Chancellor of Research **Oct 2025 - Present**
Data Engineering Domain Consultant *Berkeley, CA*
- Engineered data transfer workflows across 27 departments, accelerating ingestion by **30%** and enabling analytics and modeling with compliant cloud storage and secure computing pipelines for large-scale computation
- Automated edge-case handling workflows, scaling infrastructure while ensuring data security and privacy
- Implemented data management best practices, enhancing quality, metadata consistency, and model-readiness

Verisk Analytics **Jan 2024 - May 2025**
Data Scientist II *San Francisco, CA*
- Optimized multistage XGBoost text classification model for property claims adjustment, achieving an **0.84** F1 score using PySpark, **outperforming AWS Comprehend by 13%** and reducing production cost by **$65k/yr**
- Architected and automated a dynamic Power BI dashboard covering **60M+** property insurance claims spanning **70+** industry trends, enabling big data performance metrics visualization and cross-functional presentation
- Directed quarterly refresh of a global Regulatory Complexity Index, synthesizing data from 7 industries and **87** regression models with robust automation in ETL process, enabling business metrics evaluation
- Engineered Sentinel imagery pipeline with 6-band combinations and integrated ICLR-published Computer Vision model (**F1 0.92**) to independently pioneer geospatial analytics for insurance loss assessment
- Standardized **120+** models to ONNX format and deployed MLFlow endpoints to create low latency production
- Organized data scientist and data engineer thought leadership and teaching panel among 400+ audiences

Data Scientist I **July 2022 - Jan 2024**
- Engineered scalable election-prediction pipeline with 50K Monte Carlo simulations, optimized for distributed execution and deployed via Streamlit for real-time interactive quantitative analysis and market research
- Implemented K-Means, K-Prototypes clustering on law firms with mixed variables and achieved **84.6%** accuracy
- Spearheaded engineering of **30+** features for Workers' Comp insurance fraud risk detection with PySpark
- Extracted hierarchical data from a medical website with 5 nested layers, compiling 16K+ ICD and CPT codes

PROJECTS

Google Hackathon - 3rd Place Winner **Nov 2025 – Dec 2025**
Project Lead
- Led a 5-member team applying K-prototypes clustering to segment accounts, identifying that 0.3% of customers ("Whales") drove $2.37M in churn despite 91% of churned accounts being low-value
- Built NLP sentiment analysis pipeline quantifying 690% increase in revenue loss due to misaligned sales expectations
- Modeled data-driven retention (pre-sale alignment review) strategy projected to recover 65% of lost revenue