Feifan (Jasmine) Guan

San Francisco, CA 94015 | (609) 907-0896 | feifan.jasmine.guan@gmail.com | LinkedIn

EDUCATION

University of California, Berkeley **Aug 2025 - May 2026**
Master of Analytics
- **Coursework:** Financial Engineering, Database Design, Optimization, Risk Model & Simulation, Machine Learning

University of California, San Diego **Sep 2018 - Mar 2022**
Bachelor of Science, Data Science & Math-Economics
- **GPA:** 3.72/4.00 | **Major GPA**: 3.78/4.00
- **Coursework:** Machine Learning, Data Structures, Database Management, Statistics, Probability Theory, Vector Calculus, Linear Algebra, Econometrics, Stochastic Processes, Visualization, Scalable Analytics, Time Series

SKILLS & CERTIFICATES

- **Language & Tools:** Python, SQL, Power BI, Spark, Databricks, AWS, MLFlow, NLP, QGIS, Computer Vision
- **Analysis & Modeling**: Deep Learning, Deployment, Statistical Simulation, Financial Modeling, AB Testing
- **Certificates**: AWS Certified ML Specialty, AWS Cloud Practitioner, DeepLearning.ai Certificate, IBM DS Professional

PROFESSIONAL EXPERIENCE

Verisk Analytics **Jan 2024 - May 2025**
Data Scientist II *San Francisco, CA*
- Designed and implemented data pipeline to process Sentinel satellite imagery using 6-band combinations, integrating with **ICLR-published computer vision** model to assess property damage from civil unrest, achieving **F1 score of 0.97**
- Managed quarterly refresh of a global **Regulatory Complexity** Index, synthesizing data from 7 industries and 87 XGBoost regression models with robust automation in ETL process, and presented actionable insights
- Mapped building footprint of $1b property damaged by civil unrest event during 2020 in QGIS through satellite model
- Converted 120+ models to ONNX format and established MLFlow endpoints to create low latency deployment
- Conducted election prediction using **Monte Carlo simulation** on 50,000 samples and implemented through Streamlit

Verisk Analytics **Jul 2022 - Jan 2024**
Data Scientist I *San Francisco, CA*
- Developed multistage XGBoost **text classification** model with 4 embeddings, achieving an **0.84 F1** score using PySpark on Databricks, **outperforming AWS Comprehend by 13%** and reducing production cost by **$65k/yr**
- Engineered and maintained a dynamic Power BI dashboard covering **60M+** property claims spanning 70+ industry trends, enabling effective data visualization and communication of insights to non-technical teams
- Performed K-Means, K-Prototypes clustering on law firms with mixed variables and achieved 84.6% testing accuracy
- Led engineering of 30+ features for workers' comp fraud detection with PySpark for increased production efficiency
- Scraped medical website with 5 nested layers to compile hierarchy of 16000+ ICD and CPT codes with BeautifulSoup

Verisk Analytics **Jun 2021 - Aug 2021**
Data Science Associate Intern *Jersey City, NJ*
- Designed **7** business insight features while migrating Client Health Score dashboard from Tableau to Power BI
- Restructured 11 data relations and 6 data types in ETL process to standardize database of 10k client for adaptability
- Mapped COVID underreporting factor of **187** countries using bootstrapped linear regression on economic indices, providing insights for policy adjustments and resource allocation

PROJECTS

Jane Street Market Data Analysis **Jan 2021 - Mar 2021**
- Analyzed 10,000+ time series data and extracted 10 features of **high return and low volatility** through PCA
- Extracted optimal investment time horizon across 3 return metrics to risk adjusted estimates by **12%**
- Extrapolated volatility through concavity and convexity of quantile-quantile plot of different time horizon

Music Timbre & News Sentiment Analysis **Mar 2021 - Jun 2021**
- Utilized Python Altair heatmap and ridgeline plot for interactive visualization with linked brushing
- Conducted univariate analysis of correlation between **12** timbres and its distribution shift over 9 decades with Altair
- Analyzed correlation between music timbres and NLTK sentiment score of 6 million NYT articles

Scalable Loan Default Prediction with AWS **Mar 2021 - Jun 2021**
- Established SSH connection between VS Code and AWS S3, EC2 and EMR for cloud data pipeline
- Utilized Dask to engineer **61 mortgage features** and PySpark to extract default label of scalable observations
- Delivered engineered feature and extracted label into a Scikit-learn ML pipeline while using train-test split