Jaewon Shim

Data Analyst / Data Scientist

jaewon.shim@berkeley.edu | 510-833-0481 | Hayward, CA | www.linkedin.com/in/jae-won-shim

https://jaewonwebsite.com

EDUCATION

University of California Berkeley, *B.S.* Dec 2025 | Berkeley, CA

- Data Science (Business & Industrial Analytics) | GPA: 3.9 / 4.0

WORK EXPERIENCE

Business Analyst Intern, *DocuSign* Jun 2025 – present

- Built Tableau dashboards by integrating 50M+ Snowflake/Salesforce records, enabling leadership to monitor KPIs across functions and reducing manual reporting by 80%.
- Developed a logistic regression model in Python (92% accuracy) to predict renewal/expansion; automated scoring with Airflow and integrated into Salesforce, improving Customer Success prioritization.
- Analyzed 40K+ sales sequences to identify top outreach strategies, boosting conversion by 15% and shaping product and enablement strategy.
- Optimized complex SQL queries in Snowflake, reducing dashboard load times by 40% and enhancing usability for analysts.

Data Scientist Intern, *MKS Instruments* Jun 2024 – Dec 2024

- Designed and deployed ensemble ML models (LightGBM, Random Forest) on 500K+ inspection rows, achieving 92% accuracy, reducing false positives by 35%, and enabling real-time defect scoring that improved first-pass yield by 12% and cut scrap costs by $750K annually.
- Led migration of 20+ dashboards from Tableau to Power BI, optimizing DAX models and ETL pipelines to reduce data refresh times by 45% and improve stakeholder usability.
- Automated reporting workflows with Python and Excel VBA, boosting operational efficiency by 30% and using EDA to reduce warranty incidents by 23%.

Python and Mathematics Tutor, *Tublet* Feb 2023 – Mar 2025

- Provided tailored instruction in Python, statistics, and calculus to 100+ students, with 96% improving from C-level to A grades.
- Earned Honorable STEM Tutor Certificate (top 1%) for exceptional impact on student performance.

SKILLS

- **Programming & Data:** Python (Pandas, NumPy, Regex, PySpark, APIs), SQL (Snowflake, Datamart, Hadoop, SAP), Excel VBA, Git, dbt
- **Machine Learning & Modeling:** Scikit-learn, XGBoost, LightGBM, Causal Inference, Ensemble Methods, Clustering, Pipeline Automation, Airflow, Model Deployment, A/B Testing
- **Deep Learning:** TensorFlow, Keras, PyTorch, CNNs, RNNs, Transfer Learning
- **Statistical Analysis:** Regression, Probability, Hypothesis Testing, Quasi-Experimental Design, Linearization
- **Visualization & BI Tools:** Power BI, Tableau, Seaborn, Salesforce, Smartsheet, Outreach

PROJECTS

Defect Prediction Modeling

- Built a LightGBM-based ensemble model on 500K+ inspection records to predict final-stage laser product defects, achieving 85% recall and reducing late-stage failures.
- Automated SMOTE-based resampling and feature generation pipelines in Python, cutting preprocessing time by 40% and ensuring stable model performance across retraining cycles.
- Integrated predictions into Power BI dashboards to track defect risk trends, enabling proactive quality control and reducing scrap events by 20%.

Root Cause Analysis Dashboard

- Developed a LightGBM-based ensemble model on 500K+ inspection records to predict final-stage defects (85% recall), automated SMOTE-based resampling in Python to cut preprocessing time by 40%, and integrated outputs into Power BI to analyze OBQ, AFR, and WIRR trends—reducing scrap events by 20% and warranty incidents by 23%.

Samsung Stock Forecasting

- Implemented LSTM and GRU deep learning models in TensorFlow to forecast Samsung stock prices, achieving $R^2 = 0.95$ with the GRU model; generated 10-day predictions providing actionable insights that advised against investment due to expected price decline.

California Housing Cost Modeling

- Performed EDA and built a random forest regression model to predict housing prices, applying data preprocessing, model evaluation, and hyperparameter tuning; achieved $R^2 = 0.80$ and accurately predicted the target house price.

COVID-19 Data Exploration

- Employed PostgreSQL and advanced SQL with CTEs to perform multivariate analysis of COVID-19 data, identifying the most infectious countries and calculating a -0.751 correlation between GDP and infection rate, revealing GDP's strong inverse impact on pandemic spread.

Bike Ride Moving Average Dashboard

- Built a moving average dashboard using London bike ride data with three customizable parameters, implementing a heatmap and tooltip bar charts to visualize ride length and weather distribution for enhanced trend analysis.

CERTIFICATES

Google Data Analytics Certificate

DataCamp SQL Certificate

DataCamp Python Certificate

IBM Data Science Certificate