Pete VanBenthuysen

720-290-0999 | pete.vanbenthuysen@gmail.com | linkedin.com/in/pete-vanbenthuysen | github.com/PeteVanBenthuysen

Education

Arizona State University Tempe, AZ
Bachelor of Science in Data Science, Minor in Economics *Dec. 2025*
- GPA: 3.75

Summary

Experienced Data Scientist with a strong background in machine learning, statistical modeling, database architecture, and data quality automation. Skilled in developing and deploying predictive models, managing large-scale structured and unstructured datasets, and building scalable data pipelines that drive actionable insights and operational efficiency. Proven ability to bridge technical modeling with business and solve complex problems using innovative algorithms and optimize code for maximum efficiency. For a deeper look at my end-to-end project work, please visit my LinkedIn—or feel free to reach out directly.

Technical Skills

Languages: Python, R, SQL, JavaScript
Machine Learning & Deep Learning: Scikit-learn, XGBoost, LightGBM, PyTorch, TensorFlow, neuralnet
Statistical Tools & Methods: Time series analysis, multivariate regression, Monte Carlo simulation, Markov chains
Data Visualization: Plotly, Seaborn, Matplotlib, ggbiplot, Tableau, PowerBI
Databases: PostgreSQL, IBM DB2
Developer Tools & Platforms: Git, Docker, Kubernetes, AWS (S3, EC2, Lambda), VSCode, PyCharm, Collibra DQ
Licenses and Certifications: AWS Cloud Practitioner Candidate, Microsoft Office Specialist (Word, Excel, PowerPoint)
Leadership and Community Involvement:NLHS Society, Dean's list, ASU Devil Data Science, ASU CodeDevils

Projects

Predictive Data Quality Forecasting System | *XGBoost, Time Series, PostgreSQL, Power BI* Jun. 2025 – Jul. 2025
- Developed a scalable pipeline to forecast three months of KDE-level data quality metrics achieving average error rates under 2% per metric
- Automated ingestion and retraining workflows using a sliding window approach, uploading results to a PostgreSQL server for internal tracking and long-term storage
- Delivered forecasted outputs to stakeholders via Power BI dashboards, enabling proactive issue detection and reducing manual remediation cycles, improving operational efficiency and cost savings

Plant Disease Detection | *Python, CNN, Random Forest, Autoencoder, Grad-CAM* May 2025 – May 2025
- Developed an image-based plant disease detection system achieving 97% overall accuracy for tomato, bell pepper, and potato leaves
- Trained a convolutional autoencoder for feature compression and passed encoded vectors to Random Forest and XGBoost models for performance benchmarking
- Applied SMOTETomek resampling to correct class imbalance, boosting minority class recall by 15%+
- Built a complete evaluation pipeline with training metrics, confusion matrices, misclassification viewers, and confidence visualizations

Automated NBA Data Pipeline | *Python, Playwright, SQL, AWS, API Integration* Jan. 2025 – Present
- Built a scalable data pipeline to ingest, clean, and store NBA player and team logs, supporting over 1,000 automated daily queries across multiple seasons
- Implemented relational schema design and optimized batch inserts into SQL databases for long-term analytics
- Deployed data pipeline to AWS for cloud-based access, storage, and automated daily sync with public APIs and modular scrapers

Experience

MUFG Bank – Data Quality Solutions Tempe, AZ
Data Engineer, Summer Analyst – Financial Crimes Division *May 2025 – August 2025*
- Designed and deployed a machine learning pipeline to forecast future rule-level data quality breakage, improving operational efficiency and enabling proactive decision-making across the enterprise
- Collaborated with business teams to triage JIRA tickets and diagnose upstream/downstream data breaks across complex pipelines, reducing resolution time and improving data integrity
- Gained hands-on experience with enterprise data quality processes, including rule validation, compliance monitoring, and aligning data workflows with internal governance frameworks and regulatory standards