

Anthony Bova

ajbova333@gmail.com | 203-815-7490 | Connecticut | www.linkedin.com/in/aj-bova | Open to Relocation: Raleigh North Carolina | [GitHub: https://github.com/ajbova333](https://github.com/ajbova333)

SUMMARY

Results-driven Data Analyst and emerging AI/ML Engineer with hands-on experience in Python, statistical analysis, and machine learning who comes from a 10+ year background in sales management and business development. Adept at translating complex datasets into actionable business insights, combining analytical rigor with a proven ability to drive revenue growth, negotiate high-value contracts, and lead high-performing teams. Passionate about leveraging data-driven decision making to create measurable positive impact for organizations transitioning into AI-powered operations.

SKILLS

Python | SQL | Pandas | NumPy | Matplotlib | Exploratory Data Analysis | Statistical Analysis | Hypothesis Testing | Data Cleaning & Preprocessing | Jupyter Notebook | AWS | TensorFlow | MLflow | DVC | VS Code | Chroma DB | Hugging Face | Pytest | GitActions | Evidently | Streamlit | Nebius

TECH PROJECTS

Final Capstone Project Baseball(A)IQ APP | 9/26 | [GitHub](#)

- Designed and built a regression pipeline predicting MLB players' next-season batting average from prior performance, engineering a player-season panel dataset from the Lahman Baseball Database (1962–2023) with a time-aware train/test split to prevent data leakage
- Engineered a sabermetrically-grounded feature set (BABIP, walk/strikeout rate, age curve, positional context) to capture predictive signal beyond raw batting average alone
- Trained and compared 5 model configurations (Linear Regression, Random Forest, Gradient Boosting) across algorithm and feature-set variations, tracking experiments with MLflow and selecting the best model by MAE/R² [Tuned Linear Regression- MAE: 0.0234, RMSE: 0.0298, R²: 0.2247]
- Built a natural-language interface (Streamlit + LLM API) supporting both real-player lookups and hypothetical "what-if" scenarios, with structured input parsing, clarifying questions for missing data, and graceful handling of out-of-scope queries
- Applied production MLOps practices: config-driven training (YAML), DVC-tracked datasets, a pytest suite covering preprocessing, model validation, and interface logic

End to End MLOps Pipeline | 9/26 | [GitHub](#)

- Designed and built an end-to-end MLOps pipeline for a healthcare classification model Git, DVC, MLflow, pytest, GitHub Actions, Evidently – covering raw data ingestion through CI/CD-gated training
- Implemented DVC-based dataset versioning with an in-repository remote, enabling reproducible data access (dvc pull) from any clone with no external cloud dependencies
- Tracked and compared 5+ ML experiments across two model types in MLflow, with automated best-run selection via `mlflow.search_runs()`
- Built a 14-test pytest suite spanning unit, data validation, and model validation tiers, achieving a 100% pass rate under CI
- Configured a two-job GitHub Actions CI/CD pipeline (test → train) that automatically fails builds when model performance drops below defined thresholds
- Implemented Evidently-based drift detection comparing simulated production data against training data, generating HTML reports and enforcing a configurable drift-share threshold
- Root-caused a cross-platform data corruption bug where CRLF line-ending conversion silently broke DVC's content-addressed storage on Linux CI runner
- Diagnosed a Python import-path defect that only surfaced under the project's exact required test-invocation command

RAG-Powered Knowledge Assistant | 8/26 | [GitHub](#)

- Built an end-to-end Retrieval-Augmented Generation pipeline from scratch – document chunking, vector embeddings, semantic search, and extractive question-answering – using an 8-document, 2,100+ word knowledge base processed into 53 sentence-aware chunks with configurable overlap
- Implemented semantic retrieval with ChromaDB and Sentence-Transformer embeddings (all-MiniLM-L6-v2, 384-dim), combined with a confidence-gated response layer to distinguish answerable from out-of-scope queries
- Diagnosed and fixed a silent failure mode where out-of-scope questions returned high-confidence fabricated answers; resolved by adding an embedding-distance relevance gate, calibrated against real retrieval data, reducing false-positive confident answers from 100% to 0% across edge-case testing
- Debugged a transformers/PyTorch version incompatibility by replacing the high-level QA pipeline with a direct model inference implementation (manual logit extraction and span decoding), maintaining full functionality without downgrading dependencies
- Evaluated system performance across 15 test queries in 4 categories, identifying and documenting a structural limitation of extractive QA on comparison/multi-part questions and proposing a generative-model upgrade path

Medical Insurance Prediction Neural Network | 7/26 | [GitHub](#)

- Built and trained a deep learning regression model in Keras to predict individual medical insurance costs from demographic and lifestyle features (age, BMI, smoker status, region, etc.)
- Conducted exploratory data analysis that uncovered a significant smoker-status X BMI interaction driving cost outliers, informing feature preprocessing and model design decisions
- Diagnosed a systematic model weakness traced to function behavior on skewed target data resolved it via log-transformation, improving overall prediction accuracy by 13% and reducing relative error imbalance across the full cost range.
- Validated through controlled experimentation across multiple architectures to confirm optimal model complexity for the data size.
- Applied full ML best practices: train/test splitting prior to normalization to prevent data leakage, one-hot encoding of categorical features, and iterative model evaluation using training/validation loss curves and residual error analysis.

House Price Prediction API-FastAPI Deployment Service | 7/26 | [GitHub](#)

- Designed and deployed a production-style REST API to serve a trained machine learning model (Linear Regression, 13 features) using FastAPI, Pydantic, and Uvicorn
- Implemented model persistence and versioning (joblib serialization, semantic versioning, metadata tracking) to support reproducible model updates
- Built input validation and structured error handling using Pydantic schemas, ensuring reliable client-server data contracts
- Engineered efficient model loading (single load at startup vs. per-request) to optimize response latency
- Delivered self-documenting API endpoints (/predict, /model/info, /health) with interactive Swagger UI documentation

Beta Bank Customer Turnover Prediction | 6/26 | [GitHub](#)

- Built a binary classification model to predict customer churn from 10,000 customer records, addressing significant class imbalance (80/20 split) using scikit-learn
- Engineered and compared class-imbalance correction techniques – downsampling, upsampling, and class_weight='balanced' – across Logistic Regression, Decision Tree, and Random Forest models
- Selected and tuned a final Random Forest model (n_estimators=100, max_depth=10, class_weight='balanced', threshold=0.47), achieving F1 score of 0.607 and AUC-ROC of 0.854 on the test set
- Applied feature engineering techniques (interaction/ratio features, StandardScaler) and threshold tuning to optimize model performance beyond the required $F1 \geq 0.59$ benchmark
- Refined methodology through iterative code review, incorporating additional baseline models and resampling techniques per reviewer feedback

Megaline Mobile Plan Classification | 6/26 | [GitHub](#)

- Built a classification model to recommend optimal mobile plans ("Smart" vs. "Ultra") for 3,000+ customers based on usage behavior, using scikit-learn
- Evaluated three algorithms – Decision Tree, Random Forest, and Logistic Regression – selecting Random Forest (n_estimators=200, max_depth=7) as the final model, achieving 81.6% accuracy
- Implemented stratified train/validation/test splitting (60/20/20) to preserve class balance and prevent data leakage during hyperparameter tuning
- Refined model development through iterative code review, correcting methodology to ensure the test set was evaluated only once for unbiased final performance reporting

Venture Capital Investment Analysis | 06/26 | [GitHub](#)

Performed end-to-end SQL analysis on a comprehensive venture capital database for VentureInsight, a research firm serving VC clients making multi-million dollar investment decisions.

- Queried and analyzed a 7-table relational database spanning companies, funds, funding rounds, investments, acquisitions, and people to deliver insights for a quarterly investment report
- Engineered complex multi-table queries using INNER JOINS, subqueries, and CTEs to analyze employee education levels at failed startups and identify correlations between team background and company outcomes
- Built fund activity classification system using CASE logic to categorize venture funds into high, middle, and low activity tiers enabling clients to identify appropriate co-investment partners
- Conducted geographic and sector funding analysis using GROUP BY aggregations to identify top funded countries and US news sector investment benchmarks for international investment strategy decisions

Video Game Sales Forecasting | 05/2026 | [GitHub](#)

Analyzed historical video game sales data for a simulated online retailer to identify patterns determining game success and inform a 2017 advertising campaign strategy.

- Engineered a data pipeline in Python using Pandas to clean, preprocess, and transform a 16,700+ record dataset including handling missing values, type conversions, and TBD entries.
- Conducted platform and genre sales analysis across 12 platforms and 12 genres to identify PS4 and Xbox One as the highest priority platforms for budget allocation.
- Built regional user profiles for North America, Europe, and Japan using grouped aggregations to reveal distinct genre preferences by market.
- Executed independent samples t-tests to validate hypotheses about user rating differences across platforms and genres, providing statistically grounded marketing recommendations.

Statistical Data Analysis of Megaline Consumer Data | 04/2026 | [GitHub](#)

Analyzed usage and revenue data from 500 Megaline telecom clients to determine which prepaid plan generates more revenue and guide advertising budget decisions.

- Preprocessed and aggregated multi-table customer data in Python including rounding rules, monthly usage calculations for calls, texts, and data consumption.
- Computed statistical metrics including mean, variance, and standard deviation to compare revenue performance across Surf and Ultimate prepaid plans.
- Conducted hypothesis testing on revenue differences by plan and by region to deliver statistically supported advertising recommendations.
- Visualized key trends using Matplotlib to clearly communicate findings to non-technical stakeholders.

Exploratory Data Analysis — Instacart Customer Behavior | 03/2026 | [GitHub](#)

- Performed end-to-end EDA on Instacart's grocery orders dataset to uncover shopping patterns, peak ordering times, reorder behavior, and top products.
- Cleaned and preprocessed multiple relational tables in Pandas addressing missing values, duplicate entries, and data type inconsistencies.
- Identified peak ordering windows and top reordered products through grouped aggregations and frequency analysis.
- Designed and produced a suite of Matplotlib visualizations to communicate customer behavior insights clearly and effectively.

EXPERIENCE

Jacquet Mid Atlantic | Outside Sales Representative | 06/2024 - Present | New England and Eastern NY

Skills Used: B2B Sales, Account Management, Contract Negotiation, Revenue Growth.

- Inherited an underperforming territory at 60% of plan and drove recovery to 85% of plan by year end 2024, then grew revenue 20% year-over-year in 2025 through strategic account development and prospecting.
- Negotiated multiple long-term agreements totaling over \$1.6M in combined annual revenue across 2025 and 2026 by identifying customer needs and structuring value-based contract terms.
- Activated 17 new and dormant accounts across 2024-2026 while improving margin percentage by 4% year-over-year in 2026.

Neoperl Inc. | OEM Sales Manager | 02/2023 - 06/2024 | North America

Skills

Used: OEM Sales, Contract Negotiation, Team Management, Technical Sales

- Added 35 new customers to the portfolio in 2023 generating \$150,000 in new business revenue by prospecting across new industries and applications for existing product lines.
- Negotiated multi-year contracts tied to new customer product lines and renegotiated existing agreements to drive sustainable revenue growth.
- Managed and developed a small team of technical sales representatives through onboarding, training, performance reviews, and skills coaching.

Thyssenkrupp Materials NA | Inside Sales Manager | 05/2017 - 02/2023 | Connecticut

Skills Used: P&L

Management, Territory Management, Contract Negotiation, Solution Selling, Team Leadership

- Grew branch revenue from \$92M to \$109M over five years while managing an 11-state territory, 2 distribution hubs, and a 12-person hybrid team through disciplined KPI tracking and segment-level growth strategies.
- Negotiated defense program contracts spanning multiple years and annual agreements valued at \$1M+ in revenue with \$300,000+ in margins by leveraging solution-based and value-add selling models.
- Cultivated a high-performance branch culture through employee development, positive leadership, and structured onboarding and evaluation programs.

EDUCATION

TripleTen | AI & Machine Learning Bootcamp | 2026 - 2026

Coursework: Python programming, data cleaning and preprocessing, exploratory data analysis, statistical analysis, hypothesis testing, machine learning fundamentals, and data visualization.

Projects: Involved collaborative problem solving and iterative feedback cycles with professional reviewers simulating real-world development environments.

Southern New Hampshire University / Franklin Pierce University | Business Management & Business Administration

Coursework included business strategy, operations management, organizational behavior, and financial analysis.