Anthony Arbaiza

Hershey, PA • (570) 419-8540
Email: anthony.arbaiza124@gmail.com
LinkedIn: /in/anthonyarbaiza

EXPERIENCE

Senior Data Scientist **Sep 2023 - Present**
Natural Resources Conservation Service, USDA, Beltsville, MD

Data & Insights for Improved Hospital Operations

- **NPAD Data Refresh with R Scripting**: Initiated and led the annual NPAD data refresh process by developing and implementing R scripts. This initiative significantly improved the efficiency and accuracy of data management processes, reducing processing time by 42% and enhancing data quality. The project involved collaborating with cross-functional teams to understand data requirements and ensuring the alignment of data processing with NRCS's strategic objectives.
- **Data Suppression for USDA Sensitive Data:** Upgraded an existing R script to enhance the security and compliance of sensitive USDA data. This involved analyzing data requirements, implementing robust data suppression techniques, and conducting rigorous testing to validate script effectiveness. This effort safeguarded critical information and demonstrated my commitment to maintaining high data privacy standards.
- **CPAMS Survey Data Processing with Python:** Developed a novel Python-based solution for processing CPAMS (Conservation Practice Adoption Motivations Survey) survey data, translating intricate calculation rules into executable data processing steps. This project required a deep understanding of the survey's objectives and the application of advanced data manipulation techniques. The resulting process streamlined data analysis, facilitating more accurate and timely insights for decision-making.
- **Ad-Hoc Data Processing Requests:** Regularly managed and responded to diverse ad-hoc data requests, employing advanced statistical analysis in R. This included aligning datasets with geospatial coordinates for ArcGIS, creating data visualizations in Tableau, and developing custom solutions to meet various data analysis needs. This role underscored my versatility and ability to deliver actionable insights from complex data sets.
- **Strategic Data Management and Analysis:** Played a pivotal role in the overall data management and analysis strategy at the NRCS. My contributions have been instrumental in enhancing data-driven decision-making processes, particularly in the realms of agricultural and environmental conservation. This role involved not only technical proficiency in R and Python but also a strategic understanding of how data analytics can drive policy and operational improvements in conservation efforts.

Senior Data Scientist / Engineer **Nov 2016 - Present**
Penn State University, Hershey Medical, Hershey, PA

Data & Insights for Improved Hospital Operations

- Identified bottleneck with capacity audits, built centralized data warehouse (**PostgreSQL**) to connect separate systems (**Cerner, Powerscribe**), enabling quantification of patient care capacity.
- Managed ETL processes to ensure smooth data transition from disparate sources into our centralized PostgreSQL database, enhancing operational efficiency and data consistency.
- Spearheaded the transition from SAS to R in the initial years, optimizing data processing capabilities and ensuring a seamless transition with no operational downtime.
- Designed and built dashboards (**Grafana**) to highlight hospital overnight capacity and staffing deficiencies, reduced time to insight from 5+ days to real time and enabled radiology throughput increase by 23%.
- Developed time-series forecasting models (**ARIMA, Python, R**) to predict radiology demand and revenue, implemented custom departmental models that enabled strategic decisions on radiology center staffing.

- Collaborated with department heads to determine financial reporting requirements, built pipelines and dashboards (**Python, Pandas, SQL, Grafana, R**) to track progress on meeting revenue targets >500M$.
- Built staff-level reporting (**Grafana Dashboards, Tableau**) to highlight number of radiology exams completed and capture historic data, generated insights used to inform strategic staffing and organizational decisions.
- Implemented new data record-keeping and audit trails, built reporting and monitoring (**PGAdmin**) to ensure compliance with health privacy regulations (**HIPAA, PII**).
- Acted as technical mentor and instructor to junior analysts / researchers, taught **SQL** and **data visualization** skills to enable self-servicing of basic insight generation.

Machine Learning & Data Science to Support Patient Care

- Identified opportunity to reduce false-negative rate with CAT scans, built **classification model** (**Random Forest, SVM, Python**) to identify patient risk factors and predict progression of future diseases.
- Conducted experiments with different **machine learning models**, highlighted areas to improve patient data gathering efforts and put strategic plan in place for further data enrichment.
- Partnered with stakeholders and clinical researchers to scope and define ad-hoc analysis, conducted statistical analysis and built machine learning models to surface new insights.

Publication: Novel Use of Chatbot Technology to Educate Patients before Breast Biopsy (LINK)
Breast care is a sensitive and scary topic for patients to navigate. To ensure patients have access to resources and information regarding the details and next steps for their treatment, a chatbot was developed and provided via mobile application as an additional resource to improve patient quality of life and care.

- Conducted literature review to determine relevant clinical topics for breast biopsy and follow on treatments, queried databases (**SQL**) and developed question training dataset for chatbot development.
- Leveraged software (**AntConic, SAS**) to determine commonly used words and phrases by patients when searching for breast biopsy information.
- Conducted statistical analysis (**Fischer's Exact, Cramer's V**) to test and estimate relationship between patient question queries and relevance of recommended information.
- Created backend service to host chatbot (**Botkit.AI, Node.JS, Glitch**), deployed to online service.
- Designed user interface (HTML) mimicking text-message interfaces for use by patients on mobile devices.
- Designed decision tree chatbot architecture that enabled patients to select topics based on lifestyle, procedure info, side effects, and mental health, **87% of patients noted an improvement in sentiment**.

Database Administrator III / Architect (Top Secret Clearance) **Feb 2016 – Nov 2016**
Defense Information Systems Agency (DISA) / DDCITS, Mechanicsburg, PA

- Architected data models (**Microsoft SQL Server**) to store records for US Navy Depot.
- Established change management processes in migration from legacy database systems to **Microsoft SQL Server**, identified system dependencies and designed solutions to minimize system outages.
- Identified query bottlenecks (**SQL**) leading to decreased database performance, refactored and deployed optimized queries that reduced query runtimes by 50+%.
- Built automated **ETL pipelines** to ingest data from disparate sources into centralized warehouse.
- Utilized PowerBI to design and develop interactive dashboards and reports, providing valuable insights and facilitating data-driven decision-making processes.
- Conducted data audits and updated documentation of all data systems to ensure compliance with military data privacy regulations.
- Reviewed database error logs to determine root cause and design fixes, ensured maximum system uptime.

Database Architect (Secret Clearance) **Nov 2011 – Feb 2016**
Lockheed Martin, Air National Guard, Middletown, PA

- Conducted **end-to-end development** of new grading / academic recordkeeping features (**SQL, PHP, HTML**) to track student graduation progress, reduced time to insight from 1+ week to near real time.
- Deployed new database queries (**SQL**) and aggregations enabling population level analysis on student cohorts, surfaced insights impacting graduation rates and enabled identification of graduation roadblocks.
- Maintained SQL database (**MS SQL Server**), coordinated with stakeholders gather requirements and create ad-hoc custom queries for reporting needs.
- Developed and launched new database update and testing process, designed new unit testing to ensure accurate data, reduced data bugs that could impact student graduation rates.
- Acted as product manager, partnered with stakeholders to define new database features and requirements, constructed end-to-end project plan for development, deployment, and change management.
- Functioned as technical consultant, collaborated with end-users to identify new analytics use cases, designed and architected new technical reporting solutions to solve for the customer.

EDUCATION

Pennsylvania State University, University Park, PA **Jan 2021 – May 2023**
Masters Degree – Data Science & Applied Statistics, GPA: 3.8 / 4.0

Relevant Coursework: Applied Statistics, Data Visualization (Tableau, Python), Regression Methods & Modeling, Analytics Programming in Python, Data Mining, Large Scale Databases & Warehouses, Predictive Analytics

Pennsylvania State University, University Park, PA **Jan 2009 – May 2011**
Bachelors Degree – Information Sciences & Technology, GPA: 3.7 / 4.0

SKILLS & CERTIFICATIONS

Programming: Python (Pandas, scikit-learn, Numpy, Tensorflow, Keras, SciPy), R, SQL (MS SQL Server, PL/SQL, PostgreSQL), R, MapReduce, Apache Pig / Hadoop, Oracle, Java

Software: Tableau, PowerBI, Microsoft Word, Excel, Powerpoint, Grafana, Tableau

Others: A/B Testing, Consulting, Machine Learning (Random Forest, XGBoost, Support Vector Machine, NN, k-Means Clustering, PCA, NaiveBayes, Decision Trees), Statistical Analysis, Modeling