Mark Barranda

mark.barranda@berkeley.edu | 562 329 9725 | linkedin.com/in/markbarranda

Data Science Projects

Data Engineering / SQL Projects Summary | *SQL, PostgreSQL, MongoDB, Data Management and Querying* Fall 2024

- **Schema Exploration and Metadata Analysis**: Queried tables to analyze database structure and relationships between key tables.
- **Data Querying and Workflow Optimization**: Wrote SQL queries to explore and analyze sample tables like cast_sample, utilizing Jupyter magic commands (%%sql) to save, execute, and iterate queries efficiently, improving analysis speed and reproducibility.
- **Hierarchical Data Validation**: Designed and implemented recursive SQL queries to verify data integrity, resolving parent-child inconsistencies in complex relational datasets.
- **NoSQL Data Modeling with MongoDB**: Built and queried MongoDB collections, mastering operations such as aggregations, indexing, and schema design to handle unstructured data efficiently

Predicting Restaurant Star Ratings - Yelp | *Python, Pandas, Scikit-learn, Statsmodels* Fall 2024

- Built and compared multiple models (**Linear Regression, Regression Trees, Logistic Regression**) for predicting restaurant star ratings, optimizing for accuracy and balancing recall and false positives.
- Conducted extensive data preprocessing, including handling missing values as explicit categories and encoding categorical variables with **OneHotEncoder**, followed by removal of unnecessary variables using **p-values** and computing **Variance Inflation Factor** for multicollinearity.
- Applied cross-validation techniques using **GridSearchCV** to tune hyperparameters (e.g., ccp_alpha for tree pruning) and assessed model performance through metrics like OSR^2, MAE, and confusion matrices.

Universal Health Services, *Data Science Consultant* Summer 2024

- Drove a **33% improvement** in accuracy (from 70% to **93%**) for financial forecasting models by leveraging **Random Forest** algorithms, advanced EDA, and data-driven feature engineering across datasets from 300+ facilities.
- Independently optimized feature selection to achieve **93.35% accuracy** on Total Change in Prior Bad Debt predictions, demonstrating expertise in refining model performance through targeted filtering and parameter tuning.

The Nature Conservancy, *Data Science Consultant* Spring 2024

- Leveraged Kling Gupta Efficiency (KGE) metrics to **optimize predictive accuracy** of stream flow levels in California rivers, enhancing water resource management strategies and decision-making processes
- Led the **Exploratory Data Analysis (EDA)** on vast datasets sourced from the National Land Cover Database (NLCD), elucidating actionable insights to enhance the team's comprehension of feature utilization, thereby optimizing model accuracy across diverse applications.

World Wildlife Fund, *Data Science Consultant* Fall 2023

- Comprehensive evaluation of five leading Large Language Models to identify the most effective summarization of WWF's global accomplishments
- Guided the visualization of WWF's global impact through the integration of **Mapbox**, **Folium**, and **OpenStreetMap** polygons stored in GeoJSON formats. Led the development of an interactive map featuring a search function.

Spam vs. Ham Classification | *Python, Scikit-learn, Logistic Regression* Fall 2022

- Achieved 85.2% accuracy in classifying spam vs. ham emails by designing a logistic regression model, showcasing strong text classification and predictive modeling skills.
- Optimized performance through text vectorization, feature scaling, and detailed evaluation using a confusion matrix, identifying areas for improvement and refining results.

Ethical Evaluation of Housing Models in Cook County | *Python, EDA, Pandas, Residual Analysis, Ethics in Data Science* Fall 2022

- **Exploratory Analysis and Feature Engineering:** Analyzed over 200,000 property records, applying log transformations to address skewed distributions in sale prices, improving data symmetry and visualization. Identified Log Building Square Feet as a key predictive feature while flagging outliers for refinement.
- **Bias and Fairness Evaluation**: Investigated Cook County's property tax system, revealing regressive patterns where inexpensive homes were overvalued, disproportionately burdening non-white property owners. Evaluated predictions using Root Mean Square Error (RMSE), identifying increased errors (~1.5) for high-value properties.
- **Insights through Visualizations**: Developed visualizations to uncover socio-economic implications, such as residual distributions and overestimation percentages. Found that ~20% of high-value homes were underestimated, highlighting fairness concerns in tax assessments.

Education

University of California, Berkeley IP/ Graduation: May 2025

B.A. Data Science

Relevant Coursework: Foundations of Data Science, Principles & Techniques of Data Science, Structure and Interpretation of Computer Programs, Data Structures & Algorithms, Linear Algebra, Probability in Data Science, Geospatial Information Systems, Probability for Data Science, Data Engineering, Machine Learning and Data Analytics I (Intro), Machine Learning and Data Analytics (Advanced)

Fullerton College Graduation: May 2022

AA - Interdisciplinary Studies w/ Emphasis in Math & Science

Relevant Coursework: Introduction to Programming, Research Methods in Psychology, Brain & Behavior, Social Psychology

Volunteer & Activities

Data Science Society - Social Good Committee - Project Manager | The Berkeley Project | Intramural Volleyball Captain | GIF Berkeley

Additional Info

Skills: SQL, DBMS, CSS, Java, HTML, MongoDB, JSON, Python, predictive analytics, EDA, GIS, Scikit-learn, Numpy, geospatial analysis

Languages: English (Native), Tagalog (Native)

Certification: CITI - Social and Behavioral Research