Akul Bajaj – Data Analyst/Scientist
San Francisco, CA | akulbajaj2001@gmail.com | (916) 717-0795 | github.com/akulbajaj | linkedin.com/in/akul-bajaj

EDUCATION

University of San Francisco, California July 2022-June 2023
Master of Science - Data Science
Relevant Coursework: Data Acquisition, Computation (Python), Regression, Time Series, Machine Learning, Advanced ML, Distributed Data Systems, Data Structures and Algorithms, Statistics, Probability, Data Visualization, SQL, NoSQL, Linear Regression, Design of Experiments (A/B/n Testing), Distributed Computing (Spark), Data Analytics (PowerBI and Tableau)

UC Santa Barbara, California
Bachelor of Science - Statistics and Data Science
Relevant Coursework: Linear Regression, Machine Learning, Advanced Statistics, Time Series Analysis, Programming in SAS, Data Science for Biology

PROFESSIONAL EXPERIENCE

Metropolitan Transportation Commission, San Francisco CA November 2022-June 2023
Data Scientist

- Led end-to-end **geospatial data project** involving gathering and labeling zoning GIS links for 101 jurisdictions and 8 unincorporated regions to scrape GeoJSON zoning maps ranging from 787 KB to 30.2 MB.
  - Created functions using the GeoPandas library to map cities to their current zone codes and dissolve shape geometries for their respective zone codes.
  - Resulted in saving 15+ people on the Bay Area Spatial Information System (BASIS) team 10+ minutes of work per jurisdiction, contributing to the overarching Plan Bay Area 2050 effort to serve 7.7 million people in the Bay Area.
- Worked on **Computer Vision** project to predict land use type of google street view images.
  - Improved previous model choosing the more sophisticated Resnet18 as base model, enabling fine tuning for layers within the pretrained model, and implementing additional augmentation techniques to the training images.
  - Trained model on 8000 images and raised F1 score from .76 to .81.
- **Skills:** Github, AWS S3, AWS Redshift, arcGIS, GeoPandas, web scraping, python, Computer Vision, PyTorch, cv2, Tableau

Data Science UCSB, Santa Barbara CA September 2020-May 2022
Senior Member

- Held 4 workshops for about 30 people each time, on business analytics and model performance metrics such as R squared, MAE, MSE. Interacted with 5+ students after each event to clarify topics and answer questions
- Assisted a total of ~8 students with coursework including advanced statistics, time series analysis, and machine learning
- Presented "under the hood" methodologies behind ML methods: K means clustering, Random Forest, Decision Tree with a team of 3

Rubios Coastal Grill, Elk Grove CA February 2017-September 2021
Manager

- Managed shifts to ensure customer satisfaction, employee well-being (usually 8 employees at a time), and restaurant profit
- Trained 20+ employees on positions ranging from guest service agent (GSA), Cook, Shift Lead
- Maintained a total of 100+ surveys per month, with a satisfaction rate of 92%

U.S. Census Bureau, Sacramento CA July 2020-September 2020
Data Intern

- Recorded data from individuals regarding their households and added to the US Census data base. Interviewed approximately 10 people per day.

ACADEMIC PROJECTS

Sentiment Analysis for Amazon Reviews

- Data sources: 2 datasets, 1 static json file of 883,636 reviews for over 201,959 products, and data collected from the Amazon API.
- Machine learning: Used PyTorch and logistic regression to predict sentiment of reviews. Trained model using first 95,000 reviews. Used "bag-of-words" feature transformation and rating column as target variable. Optimized loss function using Adam and evaluated model using MAE. Model performance: MAE of 0.07339.

GDELT Analysis

- Conducted a project using GDELT data in DataBricks to analyze empathetic comments by source region. The project involved importing a massive dataset from a GCP bucket, cleaning the data using Spark and Python, and creating a scatter plot of latitudes and longitudes to visualize the results in a map. To optimize performance, caching was implemented, which allowed for faster processing of the data. The project demonstrated proficiency in Spark and big data processing, working with large datasets ranging in size up to several gigabytes.

Linear Regression Predicting Car Prices

- Fit a full model, then conducted analysis of variance, to guide feature selection. Continued with Model Diagnosis, troubleshooting observations that violate the assumptions of linear regression. Assumptions include: Heteroskedasticity, Normality of residuals and Multicollinearity. Finally, we selected a model using BIC stepwise model selection.

TECHNICAL SKILLS

HTML · Git · Linear Regression · Random Forest · Logistic Regression · k-means clustering · Machine Learning Algorithms · Data Management · Natural Language Processing (NLP) · Amazon Web Services (AWS) · Predictive Modeling · Deep Learning · NumPy · PyTorch · Pandas · Scikit-Learn · Databricks · MongoDB · Data Analytics · Data Structures · Statistical Modeling · Big Data Analytics · Google Cloud Platform (GCP) · PostgreSQL · NoSQL · Apache Spark · Data Visualization · Data Engineering · Machine Learning · Python · SQL · Data Modeling · Business Analytics · Business Development · Data Analysis · R Studio · Chat GPT · SAS · PowerBI · Tableau