Akul Bajaj

Bay Area, California | akulbajaj2001@gmail.com | +1 (916) 717-0795 | Github | LinkedIn | Portfolio | USA Citizen

EDUCATION

University of San Francisco - MS Data Science July 2022-June 2023

Relevant Coursework: Machine Learning, Linear Regression, Time Series Analysis, Programming (Python), Data Structures and Algorithms, Data Visualization, Data Analytics (PowerBI and Tableau), Data Acquisition, SQL, NoSQL, Distributed Computing (Spark), A/B Testing

UC Santa Barbara - BS 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

- Implemented end-to-end **geospatial data project** for 101 jurisdictions and 8 unincorporated regions to collect GIS zoning data
  - Created functions using the GeoPandas library to map cities to their current zone codes and dissolve shape geometries for their respective zone codes
  - Achieved time savings equivalent to 10+ minutes per jurisdiction for each member of the Bay Area Spatial Information System (BASIS) team. This contribution significantly supported the broader Plan Bay Area 2050 initiative, aimed at serving the needs of the 7.7 million residents in the Bay Area
- Worked on **Computer Vision** project to predict land use type of google street view images
  - Improved previous model using deep learning (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 by .05
- Conducted Capactiy **Data Imputation** Project, to address missing values for features like max Density Units per Acre, Floor Area Ratio
  - Extracted capacity data by formulating and executing AWS Redshift SQL queries to compile comprehensive dataset from raw data stored in an S3 bucket
  - Employed a variety of data science techniques, including GeoPandas for proximity analysis, Random Forest algorithms with feature engineering for predictive imputation, and straightforward grouping and mean calculation for effective data imputation strategies.
- Skills: Github, AWS S3, AWS Redshift, arcGIS, GeoPandas, web scraping, python, Computer Vision, PyTorch, cv2, Tableau

Data Science Club 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

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

Data Intern

- Utilized specialized data collection software and mobile applications to ensure accurate and efficient data entry, contributing to the integrity of the United States Census database
- Conducted in-depth interviews with a diverse range of respondents collecting and verifying crucial demographic, socioeconomic and geographic information

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

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