Ngoc (Alice) Hua

DATA SCIENTIST

626.677.1998
alicehua11@berkeley.edu
linkedin.com/in/alicehua/
github.com/alicehua11
alicehuacal.com

SKILLS

SQL / Python / R / PySpark / AWS / GCP/ Databricks / Git / GIS / ArcGIS / Docker / Kubernetes / Tableau / HTML / CSS / Carto / Mapbox

EDUCATION

Master of Information & Data Science
UC Berkeley
Anticipated 2021

Bachelor of Arts, Geography
Magna Cum Laude, Phi Beta Kappa
UC Berkeley. 2020

Associate of Arts,
Social Sciences & Social Behavior
Certificate, Automotive Technology
Rio Hondo College, 2018

SELECTED COURSEWORK

Applied Machine Learning
Statistics for Data Science
Fundamental of Data Engineering
Research Design
Machine Learning at Scale
Privacy Engineering

CURRENT COURSEWORK

Deep Learning in the Cloud and at the Edge
Experiments and Causal Inference

LANGUAGES

English & Vietnamese

PROJECTS

Flight Delay Prediction for ML at Scale *2021*

- Built an end-to-end pipeline for binary classification with imbalanced dataset of air traffic and weather data. Pipeline ran on Databricks with r4.xlarge cluster
- Performed ETL, established baseline model with Logistic Regression, validated final model using Gradient Boosted Trees, performed timeseries k-fold cross validation and grid search CV to find best parameters and avoid data leakage problems

Re-identification on Anonymously Reported Salaries *2021*

- Developed a proof of concept: performed join attacks via webscraped LinkedIn profiles and Zoom dataset to re-identify individually reported salaries on Levels.fyi
- Implemented Mondrian algorithm for recommended privacy protection

Regression Study of the Spread of Covid-19 *2020*

- Used classical linear model of OLS to estimate the causal relationship between adults from 19 to 34 and the number of Covid-19 cases in 50 US States
- Assessed the assumptions, estimated coefficients and confidence intervals. Established magnitude and direction that states with more young adults indeed result in higher Covid-19 cases

Machine Learning for Online News Prediction *2020*

- Scaled out Python webscraping for raw data of 2020 news articles from Forbes
- Ran multiple ML algorithms with reproduced features and accessed the errors in predicting a continuous outcome

PySpark for Business Intelligence *2020*

- Used Hadoop and Spark to build a data pipeline, from Docker cluster, consume messages from Kafka to Spark for data transformation (i.e. using flatMap() to unroll a nested json) and used PySpark SQL for querying

Descriptive Analysis: Bird Strikes & Vertiport Analysis *2019*

- Analyzed bird strikes trend from 1990 – 2018 against airline flights, temperature & migration datasets using R and LIDAR data
- Produced suitability analysis for vertiport locations in using ArcGIS for the Civil and Environment Department at UC Berkeley

HIGHLIGHTED EXPERIENCE

UC Berkeley School of Chemistry Berkeley, CA | 2021
Data Analyst

- Discover and collect large amounts of employment data of alumni using LinkedIn Sales Navigator to better engage UC Berkeley alumni and identify new donor prospects
- Analyze IPOs, recent mergers, companies that experienced unprecedented growth in recent years.

UC Berkeley Urban Displacement Project Berkeley, CA | 2019
Undergraduate Research Assistant

- Used GIS and data science skills to produce rigorous analysis on the nature of gentrification and displacement
- Used Python to wrangle and analyze urban and social media data, used ArcGIS to visualize results
- Developed interactive, online maps from the results of data analysis and develop website content
- Created Python course materials for Urban Data Analytics course to teach other students spatial and network analysis