CHAU NGUYEN

Washington D.C Metro Area | Permanent US Resident / Green card holder
chau.nng@gmail.com | github.com/chav-ngvyen | linkedin.com/in/chavngvyen | chaungvyen.com

EXPERIENCE

Sep 2021 – May 2022

Massive Data Institute Scholar at **Georgetown University**

Guided research on neural networks' limitations in object detection in paintings due to training images biases; collaborated with Art History domain expert; presented findings in research showcase

Prototyped rule-based model to identify renters in low-income areas, worked in cross-functional team of economists & engineers from 3 federal agencies to put model into production

Devised metrics to measure positive impact of allowing low-income renters to forego paperwork when applying for COVID-19 assistance

Jun 2021 - Sep 2021

Data Science Intern at **Fraym** — *geospatial ML and data analytics startup fraym.io*

Automated end-to-end log retrieval process to extract error messages from 5,200 AWS CloudWatch Logs using boto3 (Python), optimized efficiency by reducing record search and review time by 50%

Used clustering algorithm to detect text similarities in crash logs from over 1,000 failed cloud computing jobs and categorized log errors to guide engineering road-maps for next 4 quarters

Created experimentation pipeline to compare clustering algorithms which helped team select best algorithm and hyperparameter combination to put into production within 5 days

Dec 2016 – Aug 2020

Research Analyst at **International Monetary Fund**

Analyzed time-series data for over 140 million USD in overseas transfers to Samoa to measure impact of anti-money laundering compliance costs and found that prices in Pacific were 6% higher than UN targets; findings led to 4 regional conferences with 20+ countries and stakeholders

Conducted field research and utilized statistics models to estimate gaps in Tuvalu's fishing sector data accounting for over 30 million USD (55% of country's GDP)

Developed excellent verbal and written communication skills from presenting modeling decisions, analytical findings, and policy recommendations in non-technical terms to foreign government officials

PROJECTS

Identifying Original Posters from Reddit Comments (AWS, Hadoop, PySpark, SparkML, SparkSQL)

Utilized cloud computing to extract and manipulate big dataset of 8 million Reddit comments; built end-to-end gradient-boosted trees model to classify commenter's identity, achieved F-1 score of 0.92 for imbalanced class

Technical Language Processing with TVTropes corpus (NLP, gensim, NLTK, TensorBoard)

Trained custom embedding model using 40,000+ documents for plot devices which outperformed industry-standard model trained on Google News for text classification tasks in creative works

Predicting Size of Forest Fires with Convolutional Neural Network (TensorFlow, keras, sklearn)

Used CNN regression to predict the size of forest fires, performed cross-validation to prevent overfitting in small dataset; conducted feature engineering and hyper-parameters tuning, accurately predicted fire size within 0.1 km^2 margin

EDUCATION

Georgetown University | *M.S. Data Science for Public Policy* — May 2022

University of California, Berkeley | *B.A. Economics* — May 2016

SKILLS

Programming Languages: Python, R, SQL, Stata

Data Science: statistical modeling, predictive modeling, econometrics, regressions, causal inference, A/B testing, design of experiments, supervised and unsupervised ML, neural networks, NLP, data manipulation, data visualization

Technologies Used: AWS, Hive, Hadoop, SparkML, SparkSQL, PySpark, Github, Tableau, D3.js, UNIX command line