Kahang Ngau

North Bethesda, MD, 20852

ngaukahang@gmail.com | (732) 895-6216 | http://www.linkedin.com/in/kahang-ngau | Personal Portfolio

Detail-oriented data science professional with 2+ years of working experience in statistical analytics and machine learning techniques. Looking for a data scientist role to utilize my analytical and programming skills for the creation of data insights.

EDUCATION

The George Washington University, School of Engineering & Applied Science, Washington, D.C.

MASTER OF SCIENCE (M.S.) in Data Analytics, GPA 3.96/4.0   Dec 2021

- **Relevant Coursework**: *Design & Analysis of Algorithm, Applied Machine Learning, Neural Networks*

Rutgers University, School of Art & Science, New Brunswick, NJ

BACHELOR OF ARTS (B.A.) in Information Technology and Economics   May 2018

- **Dean's List**: *Aug 2016 - Dec 2017*
- **Certificates**: *IBM Data Science Professional Cert; Cert of Financial Economics; Bloomberg Market Concept Cert*

TECHNICAL SKILLS & SOFTWARE

Programming: Python, R, SQL (MySQL), Spark, Tableau, SAS, HTML, JavaScript, Streamlit, ArcGIS, and MS office.

Modeling: Linear & logistic regressions, NLP (Bag of Words), Deep learning in Neural Network (Keras and TensorFlow)

ON-CAMPUS WORK EXPERIENCE

Quantitative Research Specialist - *School of Public Health, The George Washington University*   May 2021 – Present

- Build statistical and ML classification models to predict smoke-free policy status for American public housings. Implement pipeline with normalization and prediction models, including XGBoost, Random Forest, Decision Tree, and GLM. Evaluate models' performance by using ROC curve and calculating sensitivity, specificity, and Type I & II errors in confusion metrics.
- Apply PCA and hypothesis testing to explore the NHANES survey dataset variables. Interrupt with time series on multi-linear regression model to measure the impact, using a quasi-experimental design with concept of A/B testing and causal inference.

Technical Researcher / Digital Map Developer - *GW Center for Chinese Arts and Huma*   Oct 2020 – Present

- Perform NLP in Python to search keywords and to gather digital collection links from Twitter tweets. Create interactive maps and useful browsing functions in Tableau and ArcGIS with detailed information of collections. Links: Tableau & ArcGIS
- Develop a web-based application to allow users to interact with the database of collections, along with map visualization of collections' locations. Design and implement multiple advanced searching functions in Python using Streamlit. WebApp link

WORK EXPERIENCE

Financial Data Analyst - *Bloomberg L.P., Princeton, New Jersey*   Apr 2019 – Aug 2020

- Conducted industrial research and business analysis on granular financial metrics for the top 2000 equities. Data pooling in SQL to develop internal financial data visualizations and statistical reports. Identified key trends and meaningful insights for financial decision-making. Provided technological workflow recommendations and quality assurance on client-facing data.
- Evaluated the consensus between brokers' estimated figures and companies' historical actuals in Estimate Analysis Model from Bloomberg Terminal. Built data models in Python to detect gaps and shifts between the actual and estimated figures.
- Implemented code enhancements to improve project workflow by formalizing automation in Python and Spark. Worked with the data engineering team to entail safeguarding against drift in model performance to ensure work efficiency and accuracy.

Data Reporting Analyst - *Amedisys Inc., Burlington, New Jersey*   Jun 2018 – Apr 2019

- Performed data mining and query functions across relational databases in Microsoft SQL Server. Analyzed data in R to develop strategic data reports. Utilized features and algorithms to discover areas of improvement or new opportunities with data-driven quantitative approaches. Developed metrics to measure performance and support the decision-making process.

PROJECTS

Classification on Secondhand Smoke Exposure Using Machine Learning Models   Aug – Dec 2021

- Developed and evaluated different logistic models to classify whether people are under minor SHS exposure or not, using demographic predictors. Performed EDA, data visualization, train-test split, and model iterations in R and R-Markdown. Built different statistical and machine learning models, including GLM, Random Forest, and XGBoost with hyperparameter tuning. Interpreted the model results and selected the best model with the lowest RMSE and highest AUCROC score. Project link

Prediction on YouTube Video Likes   Aug – Dec 2020

- Implemented statistical and ML regression models to predict trending YouTube video likes in Spark Databricks. Conducted data cleaning in the steps of missing value imputation, outlier removal, and autocorrelation detection. Model preparation includes data extraction, distribution & correlation analysis, NLP, and One-Hot Encoding. Built model pipeline with data transformation and k-fold cross-validation. Performed hyperparameter tuning in Decision Tree & RF models. Project link