Hongling Yang, PHD

Address : 26349 Collier Union Drive, Murrieta CA 92563
Phone : (915) 219-1391, email : hyang78227@gmail.com
LinkedIn : www.linkedin.com/in/hongling-yang-909773227, GitHub : https://github.com/hyang78227

PROFESSIONAL PROFILE

Data science professional with a PhD in statistics and more than 10 years' experience and a strong statistical and analytical background. Expertise in data mining, python, R, SAS, C++, machine learning, and statistics. Passionate about solving problems using data, and presenting insights to business audiences

TECHNICAL SKILLS

DBMS: MS SQL Server, MySQL, Postgres
Analytical Tools: SQL, Python, R, SAS.
Data Science: Data Wrangling, Data Visualization, Statistical Modeling, Predictive Analytics, Forecasting Analytics
Machine Learning: Bagging and Ensemble Methods (Random Forest, Gradient Boosting, Adaptive Boosting, etc..), Logistic Regression, SVM, Naive Bayes, Neural Network, Time Series, Recommendation Systems, Natural Language Processing, Clustering, Dimension Reduction, Pyspark , Hadoop

EDUCATION AND CERTIFICATION

***Data Science Career Track*, Springboard** **2023**
500+ hours, mentor led professional development with hands-on application in the following areas: Python, SQL, Spark, machine learning, deep learning, inferential statistics, data visualization, data storytelling

PhD in Statistic, Arizona State University **2005-2008**
Dissertation: Estimation of Additive Coefficient Models Using the Kalman Filter (C++)
Statistical Research, Stochastic Process (SAS), Data Mining, Multivariate Data Analysis (R), Mathematical Statistics

Master of Science, Statistics, University of Texas, El Paso **2003-2005**
Thesis: A Further Study of the Relationship between PM10 Level and Daily Mortality in El Paso, TX Using A Functional Linear Model (S-plus)

RELEVANT EXPERIENCE

Data Science Career Track, Springboard

Capstone Project 1: Big Mountain Ski Resort Ticket Pricing Study **2023**
GitHub - hyang78227/DataScienceG
- Applied regression analysis using Multivariate Linear Regression and Random Forest Regression, to determine a data-driven pricing strategy for Big Mountain Ski Resort, resulting in better ticket prices.
- Implemented the pipeline design pattern to streamline operations such as missing data imputation, feature scaling, regression, hyperparameter tuning and model selection, creating concise and maintainable python notebook.
- Utilized GridSearchCV for hyperparameter tuning and training models

Capstone Project 2: A Google App Store Educational Apps Rating Analysis **2023**
GitHub - hyang78227/capstone-project3
- Performed data wrangling, pre-processing and data visualization to visually and statistically explore the data and conduct explanatory data analysis
- Implemented classification analysis using Decision Tree, Random Forest and Gradient Boosting classifiers, predicting the rating class ('Low', 'High') of educational Apps
- Employed re-sampling techniques (over-sampling and under-sampling) to address class imbalance and to adjust class representations
- Utilized RandomSearchCV for hyperparameter tuning and training classifiers

Capstone Project 3: Fashion Product Image Classification with Convolution Neural Network **2023**
GitHub - hyang78227/CapstoneProjectTwo
- Utilized ImageDataGenerator to improve minority class representations
- Developed a convolution neural network (CNN) using Transfer Learning to achieve over 95% classification accuracy on fashion product images
- Implemented Hyperband optimization algorithm for hyperparameter tuning

- Employed a pre-trained CNN model (VGG16) for feature extraction and built an item-based recommendation system for fashion products

Other Projects **2023**
GitHub - hyang78227/Springboard
- Conducted a multi-class classification analysis on the South Korean COVID-19 cases and built a Random Forest Classifier to predict patient states ('isolated', 'released', 'deceased).
- Built a Light Gradient Boosting Model (GBM) to predict the presence or absence of flight departure delay for 15mins or more, using Bayesian Optimization for hyperparameter tuning and feature engineering techniques for feature extraction.
- Conduct Time Series analysis on Cowboy Cigarettes (TM, *est.* 1890) historical cigarette sale data to predict sales trend.
- Implemented K-means clustering to build a customer segmentation system on wine customers based their responses to wine offers

Statistician, School of Medicine, University of California, San Diego **2016-2017**
• Conducted explanatory data analysis on quantitative survey data from the Rakai District of Uganda Cohort Study, analyzing population-level trends of and associations between alcohol use and intimate partner violence (IPV).
• Applied structural equation modeling and prospective multiple mediation analysis to examine the temporal relationship between alcohol, IPV and HIV infection.
• Conducted a 2-arm randomized controlled experiment to test the feasibility of an integrated alcohol and IPV perpetration reduction intervention pilot program for men and to estimate effect sizes needed for future large-scale implementation of the intervention program
• Implemented retrospective cohort mixed methods (quantitative & qualitative analysis), to evaluated the impact of neighborhood-level characteristics of the built and social environment on forced sex among African American (AA) women in Baltimore area, which in turn increases risk for HIV acquisition,
• Implemented a social epidemiological, mixed methods, multilevel study design, to examined stress at multiple levels (i.e., structural, social, and physiological) to better understand the different pathways between forced sex and HIV risk behaviors .

Statistical Consultant, Department of Internal Medicine, Texas Tech Health Center **2010-2014**
• Led and guided residents in conducting medical research
• Provided research ideas and sampling designs
• Taught and helped the resident program with Statistical Computing
• Collaborating with medical doctors in clinical trials

Statistician, College of Engineering, University of Texas, El Paso **2008-2014**
Lecturer, Department of Mathematical Science, University of Texas, El Paso **2008-2016**
• Worked on geographic information system (**GIS**) research and Ride8 Project (an Ozone pollution study in El Paso, TX)
• Collaborated with other researchers and published papers.
• Taught mathematics and statistics courses

Other Certifications
• SAS Certified Advanced Programmer for SAS 9 (Certified Serial Number: AP011585v9)
• SAS Certified Base Programmer for SAS 9 (Certified Serial Number: BP038502v9)
• SAS Certified Clinical Trials Programmer Using SAS 9 (Certified Serial Number: CTP001236v9