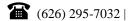# Giang Nguyen (Allison)

☎ (626) 295-7032 | ✉ gcnguyen03@gmail.com | 🌐 https://www.linkedin.com/in/giang-nguyen-794996212/

EDUCATION

California State Polytechnic University, Pomona, CA   08/2021 - 08/2023
Master of Science, Business Analytics   **GPA: 3.84 / 4.0**
Honors & Awards: Merit-Based Graduate Assistantship

Mount San Antonio College, Walnut, CA   02/2019 – 05/2021
Associate of Science, Accounting and Finance   **GPA: 4.0 / 4.0**
Honors & Awards: President's List 2019 – 2021

Duy Tan University, Da Nang, Viet Nam   08/2014 – 06/2018
Bachelor of Art, Hospitality Management   **GPA: 3.5 / 4.0**
Honors & Awards: Magna Cum Laude | Outstanding Student 2014 – 2018 | Dean List 2014 – 2018

TECHNICAL SKILLS

- **Natural Language Processing:** Proficient in NLP techniques, including working with BERT (Bidirectional Encoder Representations from Transformers) models, Named Entity Recognition (NER), Sentiment Analysis, and LSTM-based text prediction for analyzing and extracting insights from text data.
- **Data Science/ Machine Learning:** Experienced in various machine learning techniques, including Logistic Regression, Random Forest, XGBoost (Classification and Regression), Gradient Boosting, K-nearest Neighbor (KNN), K-means clustering, Linear Regression, Long Short-Term Memory (LSTM), Naïve Bayes, Optimization, A|B testing, Time Series Analysis, Neural Networks, and frameworks like scikit-learn, PyTorch, Keras, Numpy and Pandas.
- **Cloud Computing (AWS):** Familiar with AWS services such as Spark, Hadoop, S3 (Simple Storage Service), EMR, Graph Mining, Stream Mining, AWS Glue, AWS Athena, MapReduce, AWS Lambda, and AWS QuickSight for scalable and distributed data processing and analysis.
- **Visualization Tools:** Proficient in using Tableau, AWS Quick Sight, Power BI, Plotly, Matplotlib, and Seaborn for creating insightful data visualizations.
- **Programming Languages:** Skilled in Python, R, MS SQL, IBM SPSS and MATLAB for tasks related to data manipulation, analysis, and modeling.

RESEARCH EXPERIENCE

Twitter Sentiment Analytics Modeling   03/2022 – 08/2023
Research Graduate Student Assistant – Dr. Mohammad Salehan

- Conducted primary research that encompassed in-depth literature reviews, sophisticated regression methodology approaches, Sentiment Analysis, ANEW Lexicon, VADER Sentiment Intensity and the referencing of British Lexicon Project.
- Actively contributed intriguing findings with Dr. Salehan and provided seamless reports during our weekly meetings.
- Orchestrated the seamless integration of Tweepy with Twitter API V2 to create an efficient pipeline, achieving an impressive 90% reduction in the time required to collect 1.58 million tweets from six esteemed social media channels, including MSNBC, CNN, BBC World, Fox News, OANN, and Breitbart within 10 years from 2012 to 2022.
- Employed text preprocessing regard using a string-search algorithm, Regex, to lowercase characters, remove special punctuations, transform number to characters, stem words, and remove spaces from tweets.
- Analyzed emotions in tweets on social media channels by utilizing sentiment libraries, including the ANEW Lexicon, VADER Sentiment Intensity Analyzer, and NRC Lexicon, to calculate sentiment scores and analyze time responses per tweet.
- Implemented Poisson Regression by building Python scripts to regress models that examine and predict the impact of response time per character on the numbers of retweets (in natural log), offering valuable insights into the dynamics of social media engagement.
- Co-authoring a Twitter research paper (work in progress) and presented preliminary findings on "Do people react differently over time in social media" at the 2023 Western Decision Science Institute Conference held in Portland, Oregon and "The 4th CBA Showcase of Excellence" held in California State Polytechnic University, Pomona.

Imperfect Vegetable Product, Consumption Behaviors Research   08/2022 - 05/2023
Research Graduate Student Assistant – Dr. Yi Xie

- Collaborated closely with the Professor to conduct A|B testing on the Meta platform for an "Ugly Vegetable Product" advertisement, aiming to assess which variants of the advertise would perform more effectively in the competitive online market, allowing us to make data-driven marketing decisions and understand consumption behaviors.
- Conducted primary research on literature review on
- Developed Qualtrics questionnaire survey to administer 200 students to gather data insights from participants on their levels of nostalgia proneness and their attitudes related to veggie products, which are made from imperfect vegetables.
- Actively assisted the professor in collaboration with the Cal Poly Pomona Farm Store in conducting a four-weekend A|B testing, assessing customer awareness and developing hypotheses regarding whether labeling produce as "Ugly Veggie" or "Imperfect Veggie" would influence consumer interest and the likelihood of purchase.

GRADUATE DISSERTATION

Master Dissertation: Rental Properties Maintenance Data – Worked on the real project for the client, AMH (formerly known as American Homes 4 Rent). The research was conducted to provide descriptive and predictive analytics regarding maintenance needs for their inventory of rental homes nationwide. Investigated peer-reviewed research between property maintenance, age and the methodological approaches used. Developed an extensive data transformation and preprocessing phase (cleaning and handling missing values) using the Pandas framework, implementing Python to handle a substantial dataset comprising over 2.3 million records. Identified several significant factors contributing to the client's maintenance volume and costs. Imputed two normalized metrics "Work Order Count per Property-by-Property Age" and "Work Order Cost per SQFT by Property Age" to discover the relationship between property age and common property problems, geographic insights on work order trends and property distribution by state. A significant component of this dissertation revolved around the proficient development and management of a Recurrent Neural Network Long Short-Term Memory (LSTM) binary classification model. Reshaped the data frame into a 3D NumPy array as a crucial step in preparing to build an LSTM model. Initially designed to tackle plumbing issues, this model underwent meticulous coding and systematic execution. Managed the development of a LSTM binary classification model initially for Plumbing. Skillfully devised and meticulously tuned hyperparameters to make decision on Decision Threshold Trade-offs, a critical aspect of model optimization, culminating in optimized model performance and elevating precision to 79% for identifying properties in need of work orders.

Bachelor Dissertation: Improving the Quality of Hotel Banquet Staff Performance Administered a 100 guest satisfaction surveys using Google Survey Form and employed IBM's Statistical Package for the Social Sciences (SPSS) for comprehensive data analysis, including the use of statistical techniques such as Cronbach's Alpha to compute correlation coefficients among various attributes. Utilized MS Excel to create data visualizations, such as graphs and bar charts, to effectively communicate the percentage of customer satisfaction with banquet staff services. The insights gained from this analysis were instrumental in optimizing novel recruitment procedures, demonstrating proficiency in data-driven decision-making. This research significantly contributed to the enhancement of service quality within the local five-star hotel industry.

TEACHING EXPERIENCE

Marketing in Agriculture Business Management Class   08/2022 - 05/2023
Teaching Assistant | Dr. Yi. Xie, Don B. Huntley College of Agriculture
- Efficiently managed a class of 40 students, providing guidance and support for their Marketing and Advertising Study Cases projects.
- Developed comprehensive study guides, which included study problems and practice exams.
- Conducted and led weekly lecture reviews and facilitated two discussion sections for a group of 20 students.
- Assisted Dr. Xie in mentoring a team of five students participating in the National Agriculture Marketing Association competition.
- Collaborated with Dr. Xie in grading students' class activities, homeworks and assignments.

WORK EXPERIENCES

Engaging Relationships: Engagement and Philanthropic Giving   09/2022 - 05/2023
Data Consulting Intern | Collaborating with University Advancement and Singelyn Center for Innovative Analytics
- Collaborated closely with the University Advancement team and received guidance from Singelyn Center mentors, Dr. Yazdani and Dr. Kumar.
- Proficiently applied data preprocessing techniques, encompassing data cleansing, transformation, and dimension reduction, to effectively manage a substantial dataset comprising 2.2 million donor records.
- Conducted skillful analysis of alumni philanthropic activities and their engagement with the institution, successfully identifying significant giving trends.

- Played an active role in expanding the creation of milestone groups (Bronze, Silver, Gold, and Diamond), delivering actionable insights to enhance the University Advancement team's understanding of alumni donation behavior.
- Utilized a linear regression model to achieve an impressive R-squared value of 79%, enabling accurate prediction of donation amounts within the milestone groups.
- Presented research findings, titled "Engaging Relationships: Engagement and Philanthropic Giving," at the prestigious Annual Meeting of the Minds Conference in 2023.

Boeing Corrosion Text Analytics   01/2023 - 05/2023

Data Analytics Intern | Collaborating with Boeing and Singelyn Center for Innovative Analytics, CPP

- Collaborated closely with the Boeing managers and received guidance from Singelyn Center mentors, Dr. Salehan and Dr. Kumar.
- Applied a string-searching algorithm, Regex, to lowercase characters, remove special punctuations, stopwords and extract essential information from more than 20,000 free-form text records related to aircraft maintenance.
- Automatically analyzed and processed entity labeling and classified main entities with 90% accuracy into Object, Defect, and Location columns.
- Applied transformation-based learning to tagged tokens by using the chunking task format IOB (Inside, Outside, Beginning) in the realm of computational linguistics, contributing to more accurate and efficient information extraction techniques.
- Pioneered the development of a Natural Language Processing (NLP) algorithm, leveraging the state-of-the-art BERT model to achieve an impressive accuracy rate of 92.6% in detecting entities, including those containing misspelled words and demonstrating the potential for enhanced entity recognition in complex textual data.

American Home Rent Analysis of Rental Properties Maintenance   12/2022 - 08/2023

Data Analytics Intern | Collaborating with American Home Rent and Singelyn Center for Innovative Analytics

- Collaborated closely with the AMH manager and received guidance from Singelyn Center mentors, Dr. Salehan and Dr. Kumar.
- Led the development of a robust data pipeline to efficiently gather, transform, and preprocess data from a company database, involving over 2.3 million property records, including tenant information, with a specific focus on plumbing maintenance orders.
- Conducted exploratory data analysis (EDA) and cleaned the extensive property dataset using NumPy and Pandas, which included the use of lambda functions and list comprehensions.
- Collaboratively interpreted manager requirements and utilized Tableau dashboard to showcase the property age and plumbing issues, uncover geographic work order trends, and present property distribution by state to address stakeholder inquiries and requirements.
- Managed the development of a Neural Network LSTM binary classification model initially for Plumbing and, skillfully devised.
- Executed code to demonstrate decision threshold trade-offs, culminating in optimized model performance and elevating precision to 79% for identifying properties in need of work orders.

RELEVANT PROJECTS

Applying AWS (Cloud Computing) to Analyze K-12 Students Performance – Implemented a dataset of 100,000 records in an S3 bucket, establishing a robust foundation for data management. Created an efficient data pipeline in Spark to preprocess and transform the data, ensuring its readiness for in-depth analysis. Deployed Linear Regression and Gradient Boosted Tree models to identify the most significant factors affecting students' performance scores. Optimized the regression model by tuning hyperparameters, including regularization parameter (regParam) and the elastic net mixing parameter (elasticNetParam), achieving an R-squared value of approximately 93%.

Fake Reviews Detection Using Machine Learning – Extracted a dataset of 250,000 labeled records from the Yelp platform, JSON files including users, reviews, ratings and products, into Pandas dataframe. Conducted exploratory data analysis (EDA) and then applied the Scikit-learn package for data preprocessing and feature engineering before employing the Supervised Machine Learning algorithms. Conducted a comprehensive analysis, comparing the performance of Logistic Regression, Naïve Bayes, and Random Forest classification algorithms on under-sampled and over-sampled data to provide insights into effective fake review detection techniques. Notably, the Logistic Regression algorithm achieved a remarkable 90% accuracy in predicting and detecting fake reviews with under-sampled data.

Analyzing House Sale at King County – Developed an efficient data pipeline to transform and process over 200,000 records. Applied statistical analysis techniques such as hypothesis testing and quantitative analysis to validate the testing methods. Applied Pearson Correlation Matrix (Heatmap) to identify the correlation of different variables with prices and understand which variables are main factor in determining the housing price and make predictions on the

price based on the significant correlation of some variables. Applied supervised machine learning techniques, specifically linear regression, to predict house sale prices and create data visualization to observe the correlation between attributes. Applied the statistic function, Empirical Cumulative Distribution Function (ECDF) to predict the housing price with and without renovation.