DUYEN NGUYEN

duyenngu@usc.edu | 321-240-8126 | Los Angeles, CA | LinkedIn | Github | Tableau

Overview

Data Analysis, Data Manipulation, Data Visualization, Machine Learning, Statistical Analysis, Natural Language Processing

Education

- **University of Southern California** **Jan 2022 - May 2023**
  Master of Science in Applied Data Science
  - Tran Family Award (merit-based scholarship for admitted graduate students)
- **University of Central Florida** **Aug 2017 - Dec 2021**
  Bachelor of Science in Industrial Engineering, Minor in Mathematics

Technical Summary

- **Programming Languages:** Python, SQL, Matlab
- **ML Techniques:** Regression, Classification, Clustering, NLP, Image Classification, Knowledge Graph
- **ML Tools:** Scikit-learn, Pandas, Numpy, Matplotlib, Seaborn, PyTorch, HuggingFace, Tensorflow, Keras
- **Statistical Analysis:** Hypothesis testing, Regression, ANCOVA, T-test, Pearson correlation, Tukey test
- **Data Science and Miscellaneous Technologies:** Data science pipeline (cleansing, wrangling, visualization, modeling, interpretation), Microsoft Azure, Linux command line, Tableau, Time series, Excel, Git, AWS, PySpark SQL, Hadoop

Professional Experiences

- **Data Analysis and Management Coordinator - Alzheimer's MRI Data** **Apr 2022 – May 2023**
  USC Mark and Mary Stevens Neuroimaging and Informatics Institute *Los Angeles, CA*
  - Managed MRI data of over 1000 Alzheimer participants, automating a quality control pipeline, and performed data aggregation.
  - Processed MRI scan images using Python and Matlab, reducing processing time by 15% on 100,000 data points.
  - Identified patterns among Alzheimer's participant cohorts with regression models, ANCOVA, T-test, Pearson correlation, Tukey test.
- **Undergraduate Research Assistant - Data Integrity Management** **Feb 2020 – Sep 2020**
  Florida Solar Energy Center *Cocoa, FL*
  - Implemented automated email notifications using Python to monitor over 1 million daily data points, reducing 30% in data loss.
  - Identified patterns and trends through ARIMA analysis model with 2 billion of time-series data.
- **Asset and Data Management Associate** **Feb 2019 – Feb 2020**
  University of Central Florida - Facilities Operations *Orlando, FL*
  - Assisted with data standardization through internal management system AiM.
  - Attached warranty details and related documentations to of more than 2000 assets.
  - Utilized Excel to verify the accuracy and completeness of data parameters including images, locations, and group classifications.

Projects

- **SmartAlert: Event Detection through Tweets to Optimize Emergency Response Time**
  A real-time web app with AI that detects and alerts emergency departments and residents during natural disasters using tweet data
  - Tweet classification for emergency event detection using Transformer and Word2Vec Representations.
  - Achieved practical performance of ROC-AUC 0.85 via experimental validation of RF, SVM, and NN models.
  - Hosted the final classification system on the web via intergration of the ML models into Firebase app developement and visualizing disaster-impacted areas along with predicted severity and type.
  - Presented monthly progress to stakeholders and wrote final report detailing implementations taken using Lean Six Sigma methodology.
- **Vietnamese to English Machine Translation**
  A transformer-based neural machine translation model for Vietnamese to English
  - Represented word embeddings using Word2Vec, GloVe, and byte-pair encoding on 3 million English-Vietnamese sentence pairs.
  - Integrated these representations into a transformer model to improve the accuracy and fluency of Vietnamese to English translation.
- **Twitter Imaging Classification**
  An unsupervised model to detect and classify Twitter screenshots
  - Extracted features on images using pretrained VGG16 and reduced the dimensionality of the data with PCA technique.
  - Performed binary and multiclass clustering using K-means to detect and classify Twitter screenshots related to the Russia-Ukraine crisis, resulting in an accuracy of 90% and 70% respectively.
  - Presented weekly progress to mentors and peers, showcasing informative visualizations developed with Matplotlib and Seaborn.
- **Statistical Analysis of Airbnb rates in NYC**
  A forecast model to aid AirBnb hosts in setting appropriate prices for their listings while ensuring fair pricing for guests on rentals
  - Wrangled 50,000 structured data points obtained from Kaggle, built a random forest regression model to predict Airbnb rates in NYC with a result in a coefficient of determination of 0.84.
  - Showcased visualizations created with Matplotlib and highlighted influential factors determined through feature importance analysis.

Other Activities

- **Graduate Student Ambassador** **Oct 2022 - Mar 2023**
  Viterbi School of Engineering at University of Southern California
  - Assisted in answering university-life related questions to prospective and newly admitted graduate students.