LOUIS CHU

(415) 806-4691 | louischu1010@gmail.com | linkedin.com/louis-amc | louis-amc.github.io

EDUCATION

University of California, Irvine **September 2020 - June 2024**

- Bachelor of Science in Data Science, Bachelor of Arts Business Administration (Dual Degree); GPA: 3.83/4.00
- Relevant Coursework: Probability and Statistics I-III, Statistical Computing and Exploratory Data Analysis, Neural Networks and Deep Learning, Multivariate Statistical Methods

EXPERIENCE

University of California, Irvine **July 2023 - Current**

MUST Project Research Assistant *Irvine, CA*

- Performed in-depth analysis of student behavioral patterns within the Canvas Learning Management System, refining over 150 unique action patterns in student time and action counts to lay the groundwork for future machine learning investigations
- Integrated 2 distinct student term schemas within AWS Redshift, extracting 8M+ rows of student behavior data categorized by day and hourly intervals to enable data-driven decision-making, identifying most students' behavioral patterns in early afternoon
- Implemented advanced machine learning models, including Random Forest and XGBoost, to evaluate student continuity in a 3-course sequence, achieving a ROC score greater than 0.65 for specific courses using RStudio's pip operator and step function
- Delivered high-quality weekly PowerPoint presentations through Zoom, summarizing research findings for clarity, directly influencing key project decisions such as selecting predictors for fine-tuning and troubleshooting the machine learning models

University of California, Irvine **October 2022 - June 2023**

Research Assistant *Irvine, CA*

- Compiled and integrated the MIMIC-III dataset, optimizing data management workflows to achieve 50% reduction in data processing time and increased efficiency of medical data analysis using Python's pandas library and Google Cloud Platform
- Developed an advanced automation framework to extract and analyze 14,000+ medical terms to enhance clinical documentation accuracy and minimize error rates using NLTK's POS (part of speech) tagging
- Visualized and presented frequency distribution of diagnosis and prescription records to identify 20 clusters of diseases and medications pairings, guiding informed decision-making using WordCloud and K-Means Clustering

Green Dot Corporation **July 2023 - September 2023**

Data Engineering Intern *Shanghai, China*

- Tracked and managed 20+ running jobs with an average of 10+ sub-tasks, achieving success rate of 99% via Informatica's monitor service in collaboration with the database engineering team
- Performed data migration to Amazon Redshift utilizing Informatica's mapping task, improving query performance and reducing migration time in partnership with 5+ Database Engineering team members through Jira's point assignment system
- Built and optimized a machine standardized machine learning pipeline using Jupyter Notebook to analyze customer credit ratings, achieving an F1-score 0.81, and documented it on the company's Confluence page for accessibility

PROJECTS

Sepsis Data Analysis **April 2023**

- Built 7 machine learning models to analyze and predict Sepsis cases, achieving average F1 score of 0.78 and tracking losses measured by Binary Cross-Entropy through plotting lost curves, leveraging Jupyter Notebook and Python's Sklearn and NumPy libraries
- Created predictive models from patient data to enhance prediction accuracy by identifying 4 largest correlated diseases with Sepsis
- Combined table schemas with SQL relations in Azure Data Studio to better process and filter out conditions using LIKE syntax

Yelp Review Analysis **April 2023**

- Constructed a ggplot2 correlation matrix that visualized 10+ restaurant categories over a 2-day hackathon while collaborating with a team of 3, revealing a proportional relationship between ratings, waiting time, and service quality
- Initiated an NLP program by lemmatizing city names and extracting restaurant categories to classify customer reviews into time or service-related complaints, successfully creating a binary variable for enhanced analysis

CIFAR-10 Image Classification **November 2022 - December 2022**

- Conducted exploration of machine learning techniques on CIFAR-10 dataset to enhance image recognition accuracy with a team of 2
- Optimized the Random Forest model to use 20% fewer parameters, ensuring efficient usage of computational resources to achieve 48% accuracy in image classification and 50% reduction in processing time
- Performed analysis of machine learning model performance and fine-tuned over 20 hyperparameters using GridSearchCV's best params function, contributing to a notable increase in accuracy and F1-score using Python's Sklearn library

TECHNICAL SKILLS

Programming Languages: Python, R (tidyverse, ggplot2, dplyr), MySQL PostgreSQL, Amazon RedShift, C++, Java
Tools: Anaconda, Visual Studio, VMWare, FileZilla, Excel (vlookup), Azure Data Studio, PyTorch, TensorFlow, NLTK
Statistical Analysis: Hypothesis testing (t-test, f-test), Multiple Linear / Non-Linear Regression, MANOVA, GLM, NumPy, Sklearn, Pandas
Machine Learning: Convolutional Neural Networks, XGBoost, Random Forest, Logistic Regression, kNN, Clustering