Benson Duong

3234465493 | bensonduong007@gmail.com | linkedin.com/in/benson-duong-36552a180/ | benduong2001.github.io
15529 Florwood Ave. Lawndale, CA 90260 | Los Angeles Area | US Citizen

Education

University of California, San Diego Sept 2019 – June 2023
Data Science, B.S.

Experience

Business Analyst Intern June 2022 – Aug. 2022
Avanir Pharmaceuticals *Aliso Viejo, CA*

- **Extracted** from **Microsoft SQL database** and **Snowflake cloud data warehouse** with **SQL**.
- **Transformed**, cleaned the extracted SQL data with **Pandas**, **NumPy** for **feature engineering**, **data analysis**. Developed and trained **machine learning models** for predicting client engagement and sales at 78% test accuracy with **Sklearn Pipelines**. Ranked crucial business factors with SHAP feature selection and statistical testing
- Developed **GIS Python** scripts with **Geopandas** to generate data for animated, interactive **Tableau** map dashboards on region-wise business growth of sales rep territories. Presented my maps to the parent company as consulting data visualizations. Communicated my findings and recommendations to stakeholders, executives

Projects

Industry Capstone Project | *GeoPandas, Sklearn, BeautifulSoup, Git, Docker* Fall 2022 – Winter 2023

- Worked in 6-month industry research project for shipping company. Coordinated in team with Github. Developed final prediction models with 88% and 67% test accuracies that reduced costs by 9%. Showcased project to faculty and industry professionals, and authored in its research paper for company's own use.
- Procured project's geo-data and scripted its **ETL** with Socrata API, webscraping, and **GeoPandas** . Produced useful maps (i.e. delivery networks) that uncovered geographic business patterns with **clustering** methods.
- Rigorously improved ML model's accuracy from 57% to 67% through cyclical feature engineering and messy data wrangling (e.g. sample weighting) . Saved time by **automating** this iterative analysis into an end-to-end workflow python script -with ETL, **model re-training** of Sklearn pipelines, and updating of project website's github repository with generated plots and metrics auto-recorded by python logging. **Dockerized** to be reproducible.
- Taught about reinforcement learning methods leveraged by company- multi-armed bandits, Q-learning.

Restaurant Recommendation | *Recommendation ML, NLP, Unsupervised ML, Deep Learning* Winter 2023

- Built ML model on restaurant review data to predict user-restaurant recommendations with 73% test accuracy
- Performed feature engineering for review data by preprocessing review text with NLP (TF-IDF, custom-trained Word2Vec), and extracting text from review images with image-labelling neural network
- Used **Unsupervised ML** clustering to automate customer segmentation, dividing reviewers into usefully distinct, ”cuisine” sub-groups based on their review's food-related keywords

AI Image Auto-Captioner | *PyTorch, Convolutional Neural Networks, LSTM, Deep Learning* Fall 2022

- Built PyTorch neural network that generates sentence descriptions for input photos. Uses ResNet, LSTM in encoder-decoder architecture, trained for 6+ hours on university servers. Authored in its project paper, experimenting varying metrics based on hyperparameter fine-tuning

Data Engineering Sentiment Analysis on Review Text | *PySpark, Dask, NLP, AWS* Winter 2022

- Conducted data analysis and **NLP** feature engineering on **20 GB+** Amazon review text-data from **AWS S3** buckets with **Dask**, and trained ML regression models on it for predicting customer satisfaction with **PySpark**.

NYC Traffic Prediction | *ArcGIS, GeoPandas, Flask, Leaflet.js, Sklearn* Fall 2021

- Built a Python **Flask** app predicting NYC street traffic with 83% test accuracy with Sklearn **Logistic Regression** and GeoPandas for feature engineering, using clickable street map and clock input with **leaflet.js** for front-end GIS
- Authored a **Kaggle** tutorial on replicating the project using **ArcGIS** and NYC government geo-data, with matplotlib data visualizations and ANOVA hypothesis testing.

Technical Skills

Python: Pandas, NumPy, Sklearn,Geopandas, Keras, Tensorflow, PyTorch, PySpark, Dask, Flask, BeautifulSoup
SQL, NoSQL: Snowflake Data Warehouse, Microsoft SQL, DuckDB, SQLite, Neo4J
Data Visualization: Tableau, D3.js, Matplotlib, Seaborn
Others: ArcGIS, Git, Docker, R, Java, Excel,JavaScript, Leaflet.js, Selenium