Benson Duong

13234465493 | 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 Analytics Intern June 2022 – Aug. 2022
Avanir Pharmaceuticals *Aliso Viejo, CA*

- Created Python scripts with **Geo Pandas** to compile company files 2011-2021 and generate animated **Tableau** map dashboards on region-wise business growth of sales rep territory over 10 years, providing it to parent company Otsuka as consulting visuals for integration-planning. Presented maps and findings to executive stakeholders.
- Built predictive model for sales volume with 78% accuracy using Python **Sklearn** pipeline (Random Forest)
- Ranked vital business factors for client engagement with SHAP feature selection and ran statsmodel t-testing.
- Retrieved the data for these tasks by extracting from 80gb **Snowflake** cloud data warehouse and **SSMS** with **SQL** and transforming / cleaning with **Pandas, NumPy, matplotlib** for python data analysis and machine learning.

Projects

Industry Research Project | *Geo Pandas, Sklearn, git, docker, beautifulsoup* Fall 2022 – Winter 2023

- Worked in 6-month industry research team project for delivery company to optimize decision-making of business offers. Implemented sklearn regression model with 87% test accuracy, helping reduce costs by 9-10%.
- Scripted python ETL of online geo-data for project with Socrata API, **BeautifulSoup** webscraping, Geo Pandas
- Produced map visuals that uncovered useful geographic business patterns (e.g. delivery networks between zipcodes); employed K-means to cluster 33k zipcodes nodes into ”metro-regions” for locational vector-representions
- Improved ML classifier model's 57% accuracy to 67% through iterative feature engineering and messy data wrangling (e.g. sample weighting) . Automated these tasks into shell-script workflow of ETL, **dbt** with Jinja in DuckDB SQL, model re-training of Sklearn pipeline, and updating of project's Jekyll website with self-generated plots and metrics auto-recorded by Python logging in 15 minutes, with **Docker** image.
- Showcased project to faculty and industry professionals, and collaborated in project paper. Taught about reinforcement learning leveraged by company- multi-armed bandits, Q-learning.

Restaurant ML Recommender | *Recommendation ML, NLP, Tensorflow, NumPy,skLearn* Winter 2023

- Built Tensorflow neural network on text reviews, predicting user-restaurant recommendations by 73% accuracy
- Performed feature engineering for 1GB review data by preprocessing with NLP (tf-idf, custom-trained gensim word2vec), and extracting text from review images with vggnet - image-labelling Tensorflow neural network;
- Utilized unsupervised ML on user reviews with Sklearn, to automate pattern-finding for customer segmentation into separable 'cuisine' sub-groups based on reviews' distinct food-related keywords, by applying clustering on tf-idf weighted centroids of word2vec embeddings, and visualizing collaborative filtering

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

- Worked in team project for **PyTorch** neural network to auto-generate sentence descriptions for input images; programmed convolutional NN and parts of recurrent NN, trained on 18GB of 330k images for 12+ hours.

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

- Conducted data analysis on 20 GB+ Amazon review text-data from **AWS S3** buckets with **Dask**, and trained an **NLP**-feature-engineered regression model on it with **PySpark** within 45 minutes to predict customer satisfaction

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

- Developed Python **Flask** app predicting NYC street traffic by 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 with Pandas, **ArcGIS**, and 1GB of NYC government geo-data, with **seaborn** data visualizations and ANOVA, chi-square testing. Forked by 20+ users

Technical Skills

Python: Pandas, NumPy, Sklearn,Geopandas, Keras, Tensorflow, PyTorch, PySpark, Dask, Flask, BeautifulSoup
Data: SQL (PostGreSQL, MS SQL), Power BI, Snowflake, Tableau, D3.js, Matplotlib, Seaborn, Excel, R
Others: ArcGIS, QGIS, Folium, Git, Docker, Microsoft Office, Java, JavaScript, Leaflet.js, Selenium