Benson Duong

(323)-446-5493 | bensonduong007@gmail.com | linkedin.com/in/benson-duong-36552a180/ | benduong2001.github.io
Data Analysis, Data Science, Data Engineering, Business Analytics in Los Angeles Area

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*

- 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 **Snowflake** cloud data warehouse and **SSMS** with **SQL** and transforming / cleaning with **Pandas, NumPy** for python data analysis and machine learning.
- 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.

Projects

Industry Research Project | *Pandas, NumPy, Sklearn, GeoPandas, dbt, Docker* Fall 2022 – Winter 2023

- Worked in 6-month industry research team project for logistics company to optimize decision-making of business delivery offers by programming Sklearn regression model with **87% accuracy**, reducing costs **9%**.
- Scripted Python **ETL** of online geo-data for project with Socrata API, **BeautifulSoup webscraping**, GeoPandas
- Produced map visuals for **PowerBI** that uncovered subtle geographic business patterns (e.g. delivery networks between zipcodes); employed K-means to cluster 33k zipcodes into ”regions” for geographic market segmentation.
- Improved ML classifier model's **57% accuracy** to **67%** through iterative feature engineering and messy data wrangling (e.g. sample re-weighting of biased data) . **Automated** these tasks into shell workflow of ETL, pandas, **dbt** with Jinja in DuckDB SQL, model re-training of Sklearn pipeline, and updating of project's Github Repo 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 the company- multi-armed bandits, Q-learning.

Restaurant ML Recommender | *NLP, Tensorflow, NumPy* Fall 2022 - Winter 2023

- Built **Tensorflow** neural network on review data, predicting user-restaurant interaction by **73% accuracy**.
- Performed feature engineering for 1GB review data by **NLP** text-processing with **tf-idf**, custom-trained gensim **word2vec**, and text-extraction from review images with vggnet - image-labelling **Keras** neural network.
- Utilized Sklearn unsupervised ML on user review text to automate pattern-finding for customer segmentation into separable 'cuisine' sub-groups based on reviews' distinct food-related keywords, by applying clustering and **PCA dimensionality reduction** on 35k word2vec embeddings, as visual aid to **collaborative filtering**.

AI Image Auto-Captioner | *PyTorch, Deep Learning, CNN, RNN* Fall 2022

- Worked in team project for **PyTorch** neural network to auto-generate sentence descriptions for input images; implemented **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 - Fall 2022

- Developed Python **Flask** app predicting NYC street traffic by **83% accuracy** through Sklearn logistic regression and GeoPandas for feature engineering, with clock input and clickable street map via **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, Matplotlib, Geopandas, Keras, Tensorflow, PyTorch, PySpark, Flask, BeautifulSoup
Data: SQL (PostGreSQL, MS SQL), Power BI, Snowflake, Tableau, D3.js, Excel, R
Others: ArcGIS, QGIS, Folium, Git, Docker, Microsoft Office, Java, JavaScript, Leaflet.js, Selenium