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

- **Extracted** from 100GB **Microsoft SQL database** and **Snowflake cloud data warehouse** with **SQL**.
- **Transformed**, cleaned the extracted SQL data for **analysis** in Python with **Pandas**, **NumPy**.
- Built Python **machine learning** model with **Sklearn Pipelines**, predicting sales by 78% test accuracy;
- Ranked influential business factors for client engagement through statistical testing with statsmodel package.
- Created **GIS Python** scripts with **Geopandas** for generation of animated, interactive **Tableau** map dashboards on region-wise business growth of sales rep territory over 10 years. Provided maps to parent company as consulting data visualizations, and communicated findings and recommendations to executive stakeholders

Projects

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

- Worked in 6-month industry research team project for shipping company. Implemented regression model with 87% test accuracy, helping reduce costs by 9%. Showcased project to faculty and industry professionals, and collaborated in project paper. Taught about reinforcement learning leveraged by company- multi-armed bandits, Q-learning.
- Scripted python **ETL** to retrieve 1GB online geo-data for project with Socrata API and BeautifulSoup webscraping, and process it with **GeoPandas**
- Produced map visuals (i.e. delivery networks) that uncovered useful geographic business data patterns.
- Improved ML classifier model's accuracy from 57% to 67% through iterative feature engineering and messy data wrangling (e.g. sample weighting) . **Automated** these tasks into an end-to-end python and shell-script workflow with ETL, **model re-training** of Sklearn pipelines, and updating of project website with self-generated plots and metrics auto-recorded by python **logging** in 15 minutes, and **Dockerized**.

Restaurant ML Recommender | *Recommendation ML, NLP, Tensorflow, NumPy, Scikit-Learn* Winter 2023

- Built python neural network on restaurant reviews, predicting user-restaurant recommendations by 73% accuracy
- Performed feature engineering for 1GB review data by preprocessing with NLP (TF-IDF, custom-trained Word2Vec), and extracting text from review images with image-labelling Tensorflow neural network;
- Scripted **Unsupervised ML** clustering on user reviews with Sklearn, to automate pattern-finding for customer segmentation into distinguishable "cuisine" sub-groups based on reviews' distinct food-related keywords

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

- Worked in team project for PyTorch neural network that generates text descriptions for input photos; programmed CNN and parts of RNN, then trained for 6+ hours on university servers. Collaborated in its project paper.

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 for predicting customer satisfaction with **PySpark**.

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 matplotlib data visualizations and ANOVA hypothesis 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