Lawrence Lin

925-918-2478 | lawrencedlin@gmail.com | linkedin.com/in/llinnn | github.com/lawrencedlin

Education

University of San Francisco — San Francisco, CA
M.S. Data Science — *July 2021 - August 2022 (Expected)*
Courses: Advanced Machine Learning, Distributed Computing, Distributed Data Systems (NoSQL, H2o), Time Series

University of California, Santa Barbara — Santa Barbara, CA
B.S. Statistics — *August 2017 - June 2021*
Courses: Machine Learning, Bayesian Statistics, Data Structures and Algorithms, Stochastic Processes

Experience

Data Science Intern — November 2021 - Present
Walmart Labs — *Sunnyvale, CA*

- Analyzed customers' seasonal purchase behavior for festivals using Apache Spark and Seaborn
- Worked on feature engineering and data cleaning including categorical encoding and sequence creation
- Developed a Sequential Recommendation Transformer Neural Network Model in TensorFlow to recommend top-k items for a customer's next order
- Performed hyper-parameter grid-search by deploying model on a Google Dataproc cluster
- Evaluated model using AUC, Hit Rate, Mean Reciprocal Rank, and Normalized Discounted Cumulative Gain. Achieved an AUC of 0.88 and MRR of 0.56.
- Evaluated user, item, and time embedding quality by calculating the k-closest embeddings

Research Assistant — January 2021 - June 2021
Sansum Diabetes Research Institute — *Santa Barbara, CA*

- Geographically visualized diabetes severity by zip code using GeoPandas and Folium
- Performed one-way ANOVA tests and Welch's t-test and found statistically signficant differences of HbA1c with Bonferroni's correction
- Modeled HbA1c with LASSO and OLS regression models achieving an R^2 of 0.77
- Authored and presented weekly written reports of insights to SDRI researchers

Gretler Fellow — September 2019 - June 2020
University of California, Santa Barbara — *Santa Barbara, CA*

- Web-scraped thousands of pages of state legislative data using BeautifulSoup and Selenium
- Parsed raw text data using regular expressions to retrieve relevant fields
- Cleaned, processed, and validated data using Pandas

Projects

Implicit Rating Prediction) | *Pytorch, FastAi, NumPy, Pandas*

- Developed Matrix Factorization model in PyTorch
- Implemented and trained Tabular Neural Network model in FastAi using cyclical learning rate
- Created various negative sampling algorithms for models
- Achieved 1^{st} place on Kaggle leaderboard with loss of 0.4032

Search Engine |

*
*
*
*

Feature Importance | *Scikit-Learn, NumPy, Pandas*

* Implemented Spearman correlation, PCA, permutation importance, and drop-column importance
* Visualized the cross-validated Gradient-Boosted Tree's R^2 trained on k most important features
* Implemented automatic forward feature selection algorithm using permuation importance
* Calculated variance and empirical p-value of feature importances using bootstrap samples

Technical Skills

Languages: Python, R, C++, SQL (Postgres), NoSQL (Mongo) HTML/CSS, Bash
Frameworks: Hadoop Ecosystem (HDFS, YARN, Spark, SparkMLib, HiveQL) TensorFlow, PyTorch, FastAI, Scikit-Learn, Statsmodels, Scipy, Numpy, Pandas, Matplotlib, Seaborn, Flask, BeautifulSoup, Selenium, H2o
Developer Tools: Git, Docker, Google Cloud Platform, Amazon Web Services, DataBricks