Lawrence Lin

San Francisco, CA | 925-918-2478 | lawrencedlin@gmail.com | linkedin.com/in/llinnn | github.com/lawrencedlin

Education

University of San Francisco
July 2021 - August 2022 (Expected) *M.S. Data Science*
Courses: Advanced Machine Learning, Deep Learning, Relational Databases, NoSQL, Time Series Analysis, A/B Testing

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

Experience

Data Science Intern November 2021 - Present
Walmart Labs *Sunnyvale, CA*
- Analyzed millions of customers' seasonal purchase behavior using Apache Spark and Seaborn
- Worked on feature engineering and data cleaning including categorical encoding and sequence creation
- Developed a Transformer Neural Network Model in TensorFlow to sequentially recommend top-k items
- Performed hyper-parameter grid-search by deploying model on a Google Dataproc cluster and achieved an AUC of 0.88, Mean Reciprocal Rank of 0.56, and Normalized Discounted Cumulative Gain of 0.80
- Evaluated user, item, and time embedding quality by analyzing k-closest embedding by Euclidean distance

Research Assistant January 2021 - June 2021
Sansum Diabetes Research Institute *Santa Barbara, CA*
- Visualized diabetes severity by zip code using GeoPandas and Folium
- Tested for statistically significant differences in HbA1c among demographic groups using ANOVA and Welch's t-test with Bonferonni Correction
- Modeled HbA1c with LASSO and OLS regression models achieving an R^2 of 0.77

Projects

Search Engine | *Hash tables, HTML*
- Tokenized and normalized text from over seven thousand documents
- Indexed words to documents with a custom hash table implementation and displayed search results using Jinja

Implicit Rating Prediction | *Pytorch, FastAI*
- Developed a Matrix Factorization model in PyTorch and a Tabular Neural Network model in FastAI
- Trained models with negative sampling algorithms and cyclical learning rates
- Achieved 1st place on Kaggle leaderboard with a binary cross-entropy loss of 0.4032

Twitter and Reddit Sentiment Analysis | *AWS, Databricks, Spark, MongoDB, BERT*
- Extracted over a year of reddit comments and tweets mentioning an controversial celebrity using REST APIs and stored data in Amazon S3 and a MongoDB cluster
- Created new features from text using pre-trained BERT emotion and sentiment models from Hugging Face
- Predicted YouTube weekly viewership on engineered sentiment and emotion features using Random Forest and Gradient-Boosted Regression models through SparkML on Databricks cluster

Feature Importance implementation from scratch | *Scikit-Learn, NumPy, Pandas*
- Implemented Spearman correlation, Principal Components Analysis, Permutation and Drop-column importance
- Visualized the cross-validated R^2 of a Gradient-Boosted Regressor trained on k most important features
- Implemented automatic forward feature selection algorithm using permutation importance
- Calculated variance and empirical p-value of feature importances using bootstrap samples

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