Lawrence Lin

San Francisco, CA | lawrencedlin@gmail.com | linkedin.com/in/llinnn | github.com/lawrencedlin

Education

University of San Francisco July 2021 - July 2022 (Expected)
M.S. Data Science
Courses: Advanced Machine Learning, Deep Learning, Relational Databases, 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*
- Discovered peak festival shopping activity windows for millions of customers' using clustering algorithms
- Independently performed feature engineering and data cleaning on distributed datasets
- Developed and trained a Transformer Neural Network Model in TensorFlow to make personalized season-aware recommendations using historical purchases with an AUC of 0.88
- Validated time embedding quality by finding high average cosine similarities over 7-day windows

Research Assistant January 2021 - June 2021
Sansum Diabetes Research Institute *Santa Barbara, CA*
- Visualized Californian zip codes most severely impacted by diabetes using GeoPandas and Folium
- Tested for statistically significant differences in blood sugar levels among Hispanic population using ANOVA
- Modeled blood sugar levels with LASSO and OLS regression models achieving an R^2 of 0.77

Tax Intern July 2020 - August 2020
Ernst & Young *San Francisco, CA*
- Filed all tax returns for domestic and international clients ahead of schedule
- Determined compliant tax accounts by collaborating with audit teams and leveraging financial statements

Projects

Implicit Rating Prediction | *Pytorch, FastAI*
- Developed a Matrix Factorization model and a Tabular Neural Network model to predict implicit hotel ratings
- 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*
- Scraped over a year of reddit comments and tweets and stored data in Amazon S3 and a MongoDB cluster
- Engineered new features from social media with 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*
- Manually 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