Yiqun Guo

Address: New York, NY 10032 || Cell: (206) 532-6478 || ygtomor1@gmail.com
LinkedIn: linkedin.com/in/ethan-guo-7a384b369

SUMMARY

Master student who aims at finding data scientist opportunities, with sufficient project experience in business intelligence, data visualization and machine learning. Strong knowledge of statistics, data analytics and solid programming skills in Python and SQL.

EDUCATION

Columbia University New York, NY 09/2025 - 06/2027
M.S. in Biostatistics University of Washington Seattle, WA 09/2021 - 06/2025
B.A. in Geography: Data Science | Minor: Statistics GPA: 3.6/4.0

SKILLS

Programming: Python, R, SQL, Matlab
Machine Learning: Classical & Penalized Regression Methods (Lasso, Ridge), Decision Tree, Regularization, Clustering, K Nearest Neighbors, K-means, Principal Component Analysis(PCA)
Statistics Analysis: Hypothesis Testing, Experiment Design

WORK EXPERIENCE

Versuni (China) Investment Co., Ltd.
(Philips Home Appliances (China) Investment Co., Ltd.) Shanghai, CHN
Consumers and Market Insight Intern, DA Marketing Department 06/2023-08/2023

Extracted and processed sales data using Python and conducted a post-product launch review of a new Philips air fryer, based on data from the backend of two China's major e-commerce platforms, Tmall and JD.com.
Performed analyses in R (dplyr, ggplot2) to decompose air fryer sales by week, channel, and customer segment, identified high-value cohorts that accounted for 40% of revenue.
Packaged findings into deliverable visual reports (static and interactive dashboards) and handed off data products and key metrics to stakeholders to support A/B test design and pricing model development.

PROJECTS

Banking Customer Churn Prediction and Analysis
Developed algorithms for telecommunications service vendors to predict customer churn probability based on labeled data via Python programming and ApacheSpark.
Preprocessed data set by data cleaning, categorical feature transformation and standardization, etc.
Trained supervised machine learning models including Logistic Regression, Random Forest and K-Nearest Neighbors, and applied regularization with optimal parameters to overcome overfitting.
Evaluated model performance of classification (accuracy 0.86) via k-fold cross-validation technique and analyzed feature importance to identify top factors that influenced the results.

Customer Reviews Analysis and Topic Modeling
Clustered customer reviews into groups and discovered the latent semantic structures using Python.
Preprocessed review text by tokenization, stemming, removing stop words and extracted features by Term Frequency - Inverse Document Frequency (TFIDF).
Trained unsupervised learning models of K-means clustering and Latent Dirichlet Analysis. Identified latent topics and keywords of each review for clustering.
Visualized model training results by dimensionality reduction using Principal Component Analysis (PCA).

Supply Chain demand forecasting and data management
Developed a machine learning model in Python to predict future demand of different nutrition products for a pharmacy store.
Performed exploratory data analysis and preprocessed the data including handling missing values and special treatment on new product without any sales history.
Built LASSO regression and random forest models, evaluated the models via 10-fold cross validation and found optimal model parameters via grid search.
Selected the best model based on mean absolute error.