Michelle K. Li

michelle.li.524@gmail.com
michelleli.tech
github.com/michellekli
linkedin.com/in/michelleli524

Experience

Warner Bros. Entertainment Inc. — Burbank, CA
Data Scientist — 09/2019 – Present

- Pioneered programmatic method for record linkage between entities in different databases with 69% matched
- Designed and created strong user experience for data-driven web application used by core business unit
- Wrote and optimized Snowflake and Teradata SQL queries for analyses on 500 million records
- Identified and suggested solutions for data quality issues affecting multiple data sources
- Explored new libraries and created reference materials to share knowledge with team members

CGI Inc. — Los Angeles, CA
Software Engineer & Technical Consultant — 07/2016 – 01/2019

- Designed and implemented audit architecture for data analytics across mobile/backend/middleware layers
- Wrote Oracle SQL queries to identify low quality data while fixing tens of bugs
- Partnered with external developers from BAVN to design project specifications for long-term ETL process
- Interfaced with BAVN's external API to perform ETL for batch data processing every 10 minutes
- Scaled mobile application for concurrent usage across 10 locations by identifying and removing performance bottleneck
- Key team member responsible for presenting more than 5 client-facing demos

Projects

***Time Series Analysis and Forecasting for Restaurants*: (github.com/michellekli/visitor-forecasting)**

- Compared time series models (ARIMA, SARIMAX, BSTS) for forecasting daily visitors at restaurants
- Key finding: seasonal models (SARIMAX, BSTS) performed better than ARIMA
- Created pipeline to generate forecasts for any of the 800+ restaurants in the dataset

***Text Classification on Novels*: (github.com/michellekli/love-stories)**

- Compared 7 models (logistic regression, multinomial/Bernoulli/Gaussian Naive Bayes, SVM, random forest, MLP) against 5 clustering algorithms (mean shift, spectral clustering, K-means, affinity propagation, DBSCAN) for classifying novel text by author
- Key finding: multinomial Naive Bayes performed 40% better than the best spectral clustering algorithm
- Extracted features with NLP approaches: bag of words, tf-idf, word2vec, positive PMI
- Adapted techniques from research papers to estimate accuracy of clustering algorithms

Skills

Programming: R, Python, SQL, Java, C/C++, JavaScript, HTML/CSS, Git, OOP, API design, ETL scripting
Data Science: statistics, experimental design, data wrangling, exploratory data analysis, data visualization, presenting results
Machine Learning: supervised/unsupervised, regression, classification, clustering, natural language processing, time series analysis
Python Packages: numpy, pandas, matplotlib, seaborn, scikit-learn, statsmodels, SciPy, spaCy
R Packages: data.table, purrr, shiny, shinydashboard, leaflet, glmnet

Education

University of California, Los Angeles (UCLA)
BSc in Computer Science — 2016
cum laude (GPA: 3.7 / 4.0)