PAUL LIM
DATA SCIENTIST

plim0793@gmail.com
paul-lim.com
510-468-1727
San Francisco Bay Area
plim0793
plim0793

EDUCATION

University of California, Berkeley
B.S. Chemical Engineering 2016

SKILLS

Python, Machine Learning, SQL, Supervised/Unsupervised Learning, NLP, d3.js, HTML, CSS

EMPLOYMENT

METIS — San Francisco, CA
Data Science Fellow — 04/2017 - Current

- Developed five business-applicable data science projects, which utilized statistical inference, data acquisition, machine learning techniques, supervised, unsupervised learning, Flask, D3.js, and MongoDB.

RPS IRIS ENVIRONMENTAL — Oakland, CA
Staff Scientist — 08/2016 - 03/2017

- Processed laboratory analytical data to create searchable databases and concise summary tables that are compared to federal or state regulated screening levels.
- Assisted client's Environmental Health & Safety team with developing safety training courses and fitting lab spaces to be in compliance with federal regulations.

E&J GALLO WINERY — Modesto, CA
Packaging Systems Engineering Intern — 01/2015 - 07/2015

- Conducted laboratory experiments for Quality Control and Quality Assurance while applying Good Manufacturing Practices.
- Used design of experiments (DOE) to collect CO2 degradation data and fit a 2nd order polynomial regression model to the training data with an R-squared value of ~0.85. The model was used to validate the marketed shelf-life of carbonated wines.

PROJECTS

PREDICTING THE MARKET VALUE OF NBA ATHLETES

- Used the BeautifulSoup module to scrape and merge data from basketball-reference.com and ESPN.com into a single database.
- Used regression models (e.g. Linear Regression with Regularization, Random Forest Regression, and Gradient Boosting Regression) to model the data and hyper parameters for these various models were tuned using the GridSearchCV method.
- The R-squared scoring metric increased by ~0.40 after feature engineering.

CLASSIFYING HOW QUICKLY OR SLOWLY A STACKOVERFLOW QUESTION IS ANSWERED

- Scraped and compiled data from stackoverflow.com into 5 separate SQL tables stored on an AWS server.
- Used sentiment analysis and NLP to create new features that differentiated the easily answerable questions from the harder questions.
- Used classification models (e.g. Logistic Regression, Stochastic Gradient Descent, and Support Vector Machine) to predict whether a question would be answered quickly or slowly.
- The precision score was ~0.62 on the Logistic Regression model with an L2 regularization term.