Xiao (Claire) Zhang

xzhang7@scu.edu, 408-930-8189

Objective: Applied Machine Learning Engineer/Data Scientist (Full-time)

EDUCATION

Santa Clara University, Leavey School of Business (GPA: 3.5/4.0) — Santa Clara, CA
Master of Science in Information Systems — December 2018

Hangzhou Dianzi University (GPA: 4.02/5.0) — Hangzhou, China
Bachelor of Engineering in Electrical Engineering of Automation — June 2015

TECHNICAL SKILLS

- **Skills:** Python, Java, Pyspark, MySQL, Octave, Matlab, C, C++, Tableau, Pentaho
- **Libraries:** SciKit-Learn, Keras, TensorFlow, Statsmodels, Scipy, Numpy, Pandas, Matplotlib, Seaborn

WORK EXPERIENCE

KLA-Tencor | Global Service Department — Milpitas, CA
Data Scientist Summer Intern — July 2018 – Nov 2018

- Worked in a 10-people team across 2 divisions and owned 3 projects with more than 0.6-billion business impact.
- Created a data parser using python to extract specific data from html and txt log files to replace original manual data collection and boost efficiency to next level.
- Cleaned, preprocessed data and applied Cpk, Specification and correlation analysis using JMP and python on more than 10 tools' data with around 200 features; identified latent tool problems based on analysis results and increased original team efficiency and accuracy by 75%.
- Generated correlation analysis to accelerate feature engineering; came up with models including XGBoost, Random Forest, Logistic Regression to forecast future tool service action in POC process; visualized results in tableau.

ACADEMIC PROJECTS

Black Friday Purchase Prediction (Boosting & Bagging): https://github.com/clairezhang2018/Machine-Learning

- Data exploration on 550068 rows and 10 columns dataset; cross checked missing data ratio and correlation heatmap and dropped columns with high null value ratio; created 3 new features and proposed models with regard to EDA results.
- Applied XGBoost Regressor and Random Forest models; chose RMSE after cross validation as metrics. Increased model performance by around 3% after introducing 3 new features and 2 out of 3 become top5 in feature importance result of both models.

Ames House Price Prediction (Stacking): https://github.com/clairezhang2018/Machine-Learning

- Data exploration on 2920 rows and 80 columns datasets; applied bivariate and multivariate analysis to remove outliers; imputed missing data and drop columns based on null value ratio and correlation heatmap; used Box-Cox Transformations on skewed numerical data.
- Applied Lasso, Elastic Net, Gradient Boosting Regressor and XGBoost Regressor to achieve cross validation RMSE score with around 0.12, increased model performance by 6% via building up an averaged stack model with four models together.

Cat Image Recognition (Neutral Network with Classification): https://github.com/clairezhang2018/Deep-Learning

- Preprocessed data including reshaping image data into vectors and data normalization.
- Built a logistic regression model, structuring as a 2-layer shallow neutral network. After initializing parameters, defined the forward and backward propagation to learn parameters; identified cost function and computed derivatives and gradient descent to optimize model and achieved 80% accuracy.