Chia-Hui (Anita) **Shen**

Data Scientist in Analytic, Machine Learning

6363 Christie Ave Apt.1614, Emeryville, CA 94608

213-590-1148 | chshen@ucdavis.edu | aenni0409.github.io | aenni0409 | chia-hui-shen

Experienced data scientist in B2B and B2C tech companies. Demonstrated history of working experience in data-driven decision making using advanced statistics skills and building end-to-end machine learning pipeline for product development. Passionate with translating unstructured data into insights to improve product experience.

Education

M.S. in Statistics — University of California Davis, *Davis, CA* — *09/2016-12/2017*

B.B.A in Statistics — National Cheng Kung University, *Tainan, Taiwan* — *09/2011-06/2016*

B.S in Mathematics — National Cheng Kung University, *Tainan, Taiwan* — *09/2011-06/2016*

Work Experience

Machine Learning Engineer II — atSpoke, *San Francisco, CA* — *02/2020-Present*

- Applied data exploration using topic modeling with modern embedding techniques, including BERT and USE, on ticketing system text data to make product and business decisions on customer onboarding strategy.
- Scaled and automated the quarterly business reports for sales and go-to-market teams with BI tools.
- Built over 20 customer-facing dashboards and managed analytical data integrity with ETL, advanced SQL and BI tools.
- Improved and implemented knowledge-based search engine using hierarchical features with fine-tuned BERT to increase precision by 2%.

Data Analyst — Woebot Health, *San Francisco, CA* — *05/2019 -02/2020*

- Improved and evaluated over 10 new and existing product features by applying A/B testing and statistical analysis to support needs on executing product decisions with product manager and stakeholders.
- Defined user segment by exploring user retention, engagement pattern with in-depth data mining and improved user retention by 6%.
- Supported clinic researches by collaborating clinic researchers with data preprocessing, non-parametric statistical analysis, and experimental design.

Data Analyst — Metropia, Inc, *Tucson, AZ* — *04/2018-04/2019*

- Designed and implemented pattern recognition and peak-detection algorithm on time series data to discover commuters' daily travel pattern and time flexibility to change users' habitual travel schedule and increase system benefit by 20%.
- Applied real-time anomaly detection and visualized on BI dashboard to detect the abuse of incentive system.
- Built personalized incentives engine to target core users using hierarchical models with 80% accuracy.

Bioinformatics Scientist Intern — Exact Sciences Corporation (Genomic Health), *Redweed City, CA* — *06/2017-09/2017*

- Extracted and manipulated over one terabyte NGS genetic data from National Institutes of Health (NIH) by designing parallel processing system
- Developed and implemented optimization algorithms with machine learning concepts to tumor burden tracking with up to 90% patient coverage and less than 50,000 base pair panel size.

Project Experience

NCAA ML Competition 2018-Men's — Side Project, *Kaggle Competition* — *02/2018 - 03/2018*

- Applied collaborative filtering on ten-year NCAA Tourney data and player information for feature engineering.
- Implemented TrueSkill Ranking System to estimate the dynamic ranks for each possible match-ups over time.
- Predicted the outcomes of all possible match-ups by applying XGBoost with 87% accuracy.

Joke Recommendation System — Department of Computer Science, *UC Davis* — *09/2017-12/2017*

- Integrated jokes through APIs into SQL database by applying ETL process and cron job.
- Applied NLP and sensitive analysis on text from joke contents with Python and Google Cloud Platform.
- Built SVD and random forest model to predict and recommend top 10 jokes for users with 80% accuracy.

Skills

Programming Language — Python, R, Node.js, SAS

Database — RDBMS(MySQL, MSSQL, Porstage SQL), NoSQL(DynamoDB, MongoDB), BigQuery

Statistical Analysis — ANOVA, Logistic Regression, Generalized Linear Model, non-parametric statistics

Data Visualization — Tableau, Periscope, Data Studio, Plotly, Matplotlib

Machine Learning — Labeling, Random Forest, SVM, SVD, NMF, RNN, clustering, PCA

Other Tools — AWS, GCP, Serverless, CICD, Linux, Unix, Git, Matlab, Jupyter Notebooks, Excel