Shang-Chin (Jonathan), Lee

• 979-267-1233 • scleeza@tamu.edu • LinkedIn• GitHub

Python | SQL | Data Analytics | Machine learning

EDUCATION

Texas A&M University (TAMU), *College Station, Texas.* *May, 2020*
Master of Engineering in Industrial & System Engineering.
Relevant Courses: Survey of Optimization, Machine Learning and Data Analysis, Computational Tools and Database in Big Data, Simulation Modeling and Applications, Design of Experiment.

National Taiwan University (NTU) ,*Taipei, Taiwan.* *Aug, 2013*
Master of Science in Applied Mechanics.

Chang Gung University, *Taoyuan, Taiwan.* *Aug, 2011*
Bachelor of Science in Mechanical Engineering.

SKILLS

Programming Languages: SQL, Python, C++/C#, Bash/Linux, MATLAB.
Data Science Library: Pandas ,NumPy , Scikit-learn, TensorFlow/ Keras, Spacy, Matplotlib, Spark, SciPy, BeautifulSoup.
Tools: Microsoft Office, Tableau, Git.
Domain Knowledge: Statistical Test, Regression models, Supervised/Unsupervised learning, NLP, DOE (ANOVA, A/B test), Time Series Analysis, Linear and Non-Linear Programming, Inventory control, Semiconductor Physics.

EXPERIENCE

Social Impact Analytics Institute (NGO), *Seattle, WS.* *Aug 2020 – Present*
Data Scientist

- Deployed web scrappers to collect data requested by other NGO groups or text data extraction from large scale of pdf files using PyTasseract and PyMuPDF to have them into Pandas dataframe for further processing.
- Generated keywords and categories for 10Gb+ unlabeled text pdfs by unsupervised learning model, LDA, training from document corpus, using Spacy and Gensim libraries to do text cleaning and tokenization.
- Applied data mining on text files to extract information like time, location, people's names by NER tag using SpaCy library and have them as new keywords to increase web scrappers efficiency.

Chang Gung University, *Taiwan.* *Aug 2017 – Jun 2018*
Research Assistant (Full-time)

- Developed a portable ultrasound scanner prototype with convolutional kernel model to classify severe level of fatty liver by calculating its log loss using C++ (OpenCV/OpenMP), and deployed to 10+ hospitals.
- Created an GUI to store 1000+ trial testing results in a local MySQL database for further query.

Taiwan Semiconductor Manufacturing Co., Ltd. (TSMC), *Taiwan.* *Oct 2013 – Jan 2017*
Equipment Engineer (Full-time)

- Evaluated new chemical materials from different supplier, conducted experiments to find key features and established SPC chart based on statistical evidence, ensured production quality and cost effectiveness.
- Created WIP analysis spreadsheet using process log datasets to transform into pivot table and visualized the data so as to detect abnormal tools and also find window to successfully increase 6% throughput.
- Forecasted future demand of spared parts through monthly and seasonal usages and then built statistical models to predict demands and increase cost-effectiveness; saved annual expense USD 0.5M.
- Optimized supervisors' decision making streamline by generating customized KPI reports built by Microsoft Excel VBA to fast delivered data-driven insights.

PROJECTS

LSTM in Time Series Analysis

- Deployed RNN/LSTM models on a streamlit dashboard to predict covid-19 case worldwide, models were trained by using TensorFlow/ Keras library, and prediction was updated daily with CDC's latest data.

Design of Experiments

- Advanced the performance of a helicopter prototype by 20% , using design matrix, ANOVA, and OLS model to build high resolution experiments, statistically prove feature importance, and predict optimum design.

Customer Questionnaire Analysis

- Analyzed datasets from 1000+ customers questionnaires, using Scikit-learn library to do KNN imputation to fill MAR missing values with observations have similar behavior patterns.
- Trained models to predict whether people would stay after free trials, using Scikit-learn to do grid search with multiple models and found feature importance by random forest and linear regression model.

Team Mate Recommendation

- Created a team mate suggestion function for TAMU datathon website, using Scikit-learn library to preprocess 500+ questionnaires datasets and then clustering similar participants by K-mode algorithm