Xuanyi Zhao

fidelzhao218@gmail.com | 267-637-6265 | Santa Clara, CA | linkedin.com/in/xuanyi-fidel-zhao | github.com/XuanyiZhao

Education

MS in Data Science - University of Pennsylvania **Aug. 2018 - May 2020**

- ✧ **Selected Courses:** Algorithms, Big Data Analytics, Machine Learning, Database, Deep Learning, Time Series, Optimization
- ✧ **GPA:** 3.97 / 4.00

BA in Financial Engineering - Xiamen University **Aug. 2014 - June 2018**

- ✧ **Selected Courses:** Linear Algebra, Probability, Stochastic Process, Econometrics, Financial Engineering, Risk Management
- ✧ **GPA:** 3.60 / 4.00

Experience

Research Assistant **Feb. 2019 - Present**

University of Pennsylvania, Computer and Information Science Department Philadelphia, PA

- ✧ Used mixed-integer optimization to implement the optimal classification trees algorithm created by MIT researchers, improved F1-score upon traditional classification trees by 3% given 50 test datasets (Python, Gurobi)
- ✧ Built data pipelines on MIMIC-III database to extract and format 20 GB clinical data for sepsis treatment analysis (SQL, GCP)
- ✧ Applied transfer learning with group sparsity on sub-domain wiki word embeddings to explain semantic biases of polysemy

Data Scientist Intern **May 2019 - Aug. 2019**

Tencent - *the largest social media company in China* Shenzhen, China

- ✧ Led the Deepnet project, developed machine learning pipelines to identify anomalous E-commerce users based on security criteria, constructed ETL tasks and implemented NLP algorithms over 100 TB data (SQL, PySpark)
- ✧ Created word embedding and item embedding to find similar anomalous words and Apps with 85% accuracy
- ✧ Applied MinHashLSH on co-occurrence matrix and DeepWalk on graph embedding, found out over 10,000 abnormal user IDs
- ✧ Distinguished polysemous words, improved group embedding and found 2,000 abnormal chatting groups (TensorFlow)

Data Analyst Intern **Mar. 2018 - June 2018**

Vanke - *Fortune Global 500 real estate company with $45 Billion market cap* Xiamen, China

- ✧ Developed a web scraping tool to collect business and geospatial data on over 3,000 companies (Python, Scrapy, Selenium)
- ✧ Constructed an interactive dashboard on industry segmentation insights and distribution's change over time (SQL, Tableau)
- ✧ Optimized building investment decisions by predicting customer's space demand using customer segmentation (K-means)

Data Analyst Intern **Sept. 2017 - Dec. 2017**

KPMG Beijing, China

- ✧ Assisted China Construction Bank to deploy credit risk measurement engine, conducted metrics visualization (SQL, Tableau)
- ✧ Generated queries to test risk measurement engine and the integrity of CCB derivatives databases over 500 GB financial data

Projects

E-commerce Customer Segmentation and Retention Prediction (3rd place out of 35 teams in Wharton Datathon)

- ✧ Segmented customers by monthly revenue contribution using classification models on 10 GB transaction data with recall 88%
- ✧ Predicted customers' next purchasing time using MLP with recall 85%, made recommendations based on feature ranking

Procurement Anomaly Detection [*Python, Scikit-learn, PyTorch, AWS*]

- ✧ Built models to identify high-risk procurement behavior based on risk criteria over recent 4 years' UPenn purchasing data
- ✧ Explained anomaly reasons with LOF results, used Adversarial Autoencoder Neural Networks to detect latent space anomalies

LEGO Bricks Detection [*Python, Autodesk Maya, PyTorch, AWS*]

- ✧ Wrote scripts to automatically generated 4 GB single Lego bricks pictures and 1 GB assembled LEGO models' pictures
- ✧ Implemented transfer learning on ResNet-18 to distinguish 50 different single LEGO bricks with test accuracy 97%
- ✧ Conducted bricks detection on assembled LEGO models by Tiny YOLO-v3 with 62% IoU and YOLO-v3 with 73% IoU

Airbnb House Pricing Tiers Prediction [*Python, Scikit-learn*]

- ✧ Predicted prices and house tiers of new houses in Great Los Angeles Area based on 4 GB housing conditions data
- ✧ Extracted location characteristics features, applied a stack of tree-based models on house tiers prediction with 87% accuracy

Technical Skills

- ✧ Languages: Python, SQL, MATLAB, R, JavaScript, HTML
- ✧ Database: MySQL, PostgreSQL, MS SQL Server, MongoDB, HDFS, Neo4j
- ✧ Techniques & Tools: PyTorch, Spark, Hive, TensorFlow, Tableau, AWS(S3, EC2, RDS), GCP, Git, Docker