Xuhesheng Chen

4706597805 | chxuhesh@gmail.com | Location: OC, MI | linkedin.com/in/XuheshengChen | https://chxuhesh.github.io/

EDUCATION

University of North Carolina at Chapel Hill **NC, US**

Master of Information Science, GPA: 3.8/4.0 *Sep.2021 - May.2023*

- Core Course: Deep Learning, Machine Learning, Natural Language Processing, Information Retrieval, Data Mining

East China Normal University **Shanghai, China**

Bachelor of Engineering in Software Engineering *Sep.2010 - Jun.2014*

- Main Course: Probability Theory and Statistics, Database, Java Programming, Data Structure and Algorithms

Research Interests: Deep Learning, Data Engineering, Software Engineering, Artificial Intelligence.
with Agile software development, DevOps best practices, and CI/CD pipelines

PROFESSIONAL EXPERIENCES

eMatrix Energy Systems. Inc **MI, US**

Data Engineer *May.2022 -*

- Boosted 300% runtime efficiency, saved 140% cost by constructing automated data ETL pipeline with hands-on coding, aggregated data sources from different hardware platforms into Odoo ERP environment.
- Designed and created data schemas for battery manufacturing, R&D, and inventory departments, developed multiple data applications across different teams in Python (Tkinter, Flask, Apache Airflow) with DevOps practices and CI/CD pipelines (GitLab).
- Designed, developed, and managed cloud immigration process to Google Cloud Platform (GCS, Dataflow, BigQuery, Cloud Composer, Pub/Sub, Looker) for company's data warehouse.
- Administer and maintain IT infrastructure security (Google Workspace, Google Cloud, Microsoft 365, Odoo)

Ernst & Young Parthenon **Shanghai, China**

Business Analyst *Feb.2021 - Jun.2021*

- Utilized Python to collected and preprocessed 20M+ rows of multi-source datasets towards oil industrial development in last 20 years and built linear model through scikit-learn with 50+ features such as oil price, basic cost (materials, labor) and the city's GDP to predict the industry revenue in the coming 5 years.
- Provided insight about labor loss impact on fiscal decrease based on scholar papers and models, delivered solutions aiming to boost retention by 16% in next 3 years from perspectives of work-study combination and disciplines upgrading in universities.
- Visualized oil industrial distribution and revenue change with Tableau and presented insights about transferring the main industry in that city from oil to tech R&D industry.

Capgemini **Shanghai, China**

Business Analyst *Aug.2020 - Oct.2020*

- Applied DBSCAN and k-means clustering on 723 survey samples to identify demographic segmented consumers' preferences for online healthcare consulting by considering region, age, online utility time, monthly revenue as major features in Python.
- Led 3 members to conduct research on business models and market trends of current Chinese healthcare industry and presented results to investors to propose investment strategies on 5 major business tracks, including online diagonals, upgrading medicine flows and online medical insurance market expension etc, which will breakthrough 10 trillion dollars of revenue in 2023.

PROJECTS

Fake Reviews Detection of Yelp *Sep.2021 – Nov.2021*

- Specified fake words pattern with Mutual Information, Chi-Square measures and Jaccard similarity as baseline model.
- Tried with classifying typical machine-generated text based on sentimental analysis score, feature engineering with word2vec.
- Split and sampled datasets based on reviewer id to avoid information leak in valid and test sets.
- Trained Tree Models (Scikit-learn), CNNs, LSTMs, fine-tuned BERT (Pytorch) model to improve performance, fine-tuned BERT achieves precision 77%.

Wells Fargo Fraud Detection *Nov.2021 – Dec.2021*

- Performed statistical modeling on 14k+ customers transactions using PrecisionAtRecall80 to capture fraud transactions patterns and create advanced fraud prevention mechanisms to reduce risk loss for elder customers.
- Conducted stratified random sampling to solve imbalanced datasets, keeping label ration in each dataset to avoid analytics bias impact on business judgement.
- Set up strategies about model selection based on PCA analysis, tunned hyperparameter with F1 score, and achieved 95.65% accuracy on Random Forest, 96.7% accuracy on Adaboost model. Adaboost got lowest false negative rate.

Trading Factor: Shake Shack Sales Prediction from MTA Data *Jun.2021 – Aug.2021*

- Applied Python (BeautifulSoap4, geopy) to collect data from MTA and Shake Shack websites to dig novel insights about passenger traffic in New York related to Shake Shack Sales trend from 2015 to 2021.
- Preprocessed and transformed 8M+ raw datasets, and generated new features based on further data exploration and business comprehension from quarterly financial reports of Shake Shack.
- Used the prediction result as trading factor to trade Shake Shack stock

SKILLS

- **Programming:** Python (Pytorch, Tensorflow, Pandas, NumPy, Sk-learn, Beautifulsoup4), SQL, R, Tableau, Javascript, Java, C++
- **Models**: Regression (Ridge, Lasso), Logistic, SVM, Ensemble Learning, Clustering; DRL, CNN, RNN, LSTM, BERT
- **Certification**: Google Cloud Professional Data Engineer, Tableau Desktop