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

PROFESSIONAL EXPERIENCES

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

Data Engineer *May.2022 -*

- Boosted runtime efficiency by 300% and saved 140% in costs by constructing automated data ETL pipelines. Aggregated data from diverse hardware platforms into a SaaS environment using XMLRPC, ensuring seamless data integration and processing.
- Designed and developed data schemas for battery manufacturing, R&D, and inventory departments. Built multiple data applications using Python frameworks such as Flask and FastAPI, and orchestrated workflows with Apache Airflow.
- Led the cloud migration process to Google Cloud Platform (GCP), utilizing GCS, Dataflow, BigQuery, Cloud Composer, Pub/Sub, and Looker. Managed the company's data warehouse, ensuring robust and scalable data storage and analytics capabilities.
- Administered and maintained IT security infrastructure, including Google Workspace, Google Cloud IAM Policy, Microsoft 365, and Odoo. Ensured data security and compliance with industry standards.
- Collaborated with cross-functional teams to deliver high-quality data solutions. Employed version control with Git, automated testing, and implemented DevOps practices and CI/CD pipelines with GitLab, enhancing development efficiency and code quality.

Ernst & Young Parthenon **Shanghai, China**

Business Analyst *Feb.2021 - Jun.2021*

- Leveraged Python (Pandas, Numpy), SQL, and Apache Spark to collect and preprocess over 20 million rows of multi-source datasets related to the oil industry's development over the past 20 years. Ensured data integrity and consistency for robust analysis.
- Developed predictive models using scikit-learn, XGBoost, and TensorFlow, incorporating over 50 features, including oil prices, material, and labor costs, and city GDP to forecast industry revenue for the next 5 years.
- Utilized Matplotlib and Seaborn to create visualizations of oil industry distribution and revenue trends with key insights and recommendations to stakeholders, facilitating the strategic transition from an oil-based industry to a tech R&D-focused economy.

Capgemini **Shanghai, China**

Business Analyst *Aug.2020 - Oct.2020*

- Applied advanced clustering algorithms such as DBSCAN and k-means using Scikit-learn and MLlib to segment demographic data and consumer preferences. Built robust data architectures to support complex machine learning and analytical applications.
- Utilized NLP techniques with SpaCy to extract insights from expert interviews and focused on upgrading nationwide and regional Hospital Management Information Systems, with a total projected investment of 60 billion dollars over 3 years.
- Utilized Tableau to create interactive dashboards and visualizations. Communicated complex data insights effectively to stakeholders, enhancing strategic decision-making processes.

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.
- 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 a trading factor for stock.

SKILLS

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