Zoe Chen

929-996-3230 | zoechanry@gmail.com | linkedin.com/in/zoe-chen | https://github.com/zoechenry

EDUCATION

Columbia University | M.S. Data Science — *New York, NY* — 09/2022 - 12/2023
Peking University | B.S. Software Engineering — *Beijing, CN* — 09/2020 - 07/2022
Shanghai University | B.S. Finance — *Shanghai, CN* — 09/2016 - 07/2020

TECHNICAL SKILLS

Programming: Python (Pandas, NumPy, Scrapy, Scikit-Learn, TensorFlow, PyTorch), R(dplyr，ggplot2), SQL, Java
Tools: PySpark, MySQL, ETL, AWS, GCP, Tableau, PowerBI, Amplitude, GitHub
Methodologies: Machine Learning (RF, SVM, XGBoost, KNN, KMeans, Classification, Regression), Deep Learning (Neutral Network, LSTM), NLP (Transformer, Bert, GPT, Large Language Model), Hypothesis Testing, A/B Testing, PCA

INTERNSHIP EXPERIENCE

2+ years of industrial experience in leading finance and tech companies

AI Data Engineer — 06/2023 - Present
Oortech (web3 Data Cloud) — *New York, NY*

- **Deployed multimodal Large Language Models** (MiniGPT4 -7B, MiniGPT4 -13B, Visual GLM-6B) remotely, **assessed** their **performance** in perception & cognition tasks using the MME dataset, summarizing common issues.
- Designed a **distributed web crawler** based on the **Scrapy-Redis** framework to scrape 10k+ NYC food data points from Yelp, customizing an AI chatbot to serve as a NYC food guide.
- Applied **OCR** to perform **text recognition** on images provided by Lenovo, extracting 1k+ Q&A data related to Lenovo printers.

Data Scientist — 03/2022 - 08/2022
OKX/OKCoin (Global Top 10 Crypto Exchange) — *Remote*

- Applied feature engineering and built an **XGBoost** model to predict user **Life Time Value**, yielding 9.24 RMSE.
- Used **LT*ARPU** to estimate LTV in bull/bear/neutral markets, cross-validating results based on **Retention Rate & Churn Rate**.
- Developed **40+ Amplitude dashboards** for transactions & users, detecting abnormal data and designing data monitor reports.
- Designed **ROI metrics & automated report pipelines,** improving ads ROI by 15% via user conversion funnel analysis.
- **Optimized query speed** by 7% through cleaning data and creating temporary tables.

Data Scientist — 05/2021- 10/2021
Meituan (Biggest Food Delivery Platform in China) — *Beijing, CN*

- Extracted data and created temporary tables using **SQL**, applied data cleaning, selected **30+ features**, labeled training set base on AI calls and constructed an **SVM** model, accurately **identifying 85% of the 5M** wrongly labeled restaurants.
- Performed second-level stratified sampling, estimating the entire dinner cooperation space across China with statistical methods.
- Cleaned & analyzed 3M order records with **Pandas & NumPy**, identifying why **2M** good restaurants offer disappointing deals.

Data Scientist — 03/2019 - 12/2019
Ford Motor (Car-hailing Department) — *Shanghai, CN*

- Designed **5+ A/B tests** to **optimize coupon distribution** (amounts/ usage times/ validity periods) for enhanced user retention, analyzed results using **PowerBI,** proposed strategies, enhancing **3 product features and overall user experience**.
- Collected data using **SQL** and classified users via FRM & Pyramid, generated order heat maps with **Seaborn,** and proposed strategies, resulting in a 20% increase in **user retention** and a 50% decrease in **customer acquisition cost**.

PROJECTS

Smart Contract Vulnerability Detection Model | BeautifulSoup, BFS, DFS, Transformer, Bert — 01/2023 - 03/2023

- Used BeautifulSoup to scrape Smart Contracts, applying solidity-parser to generate Abstract Syntax Trees (AST).
- Applied BFS & DFS on the AST, transforming them into structured string sequences as input.
- Augmented the BERT vocabulary, training a multi-label classification model based on BERT.
- Successfully predicted vulnerability multi-label in Smart Contracts, achieving a Hamming Score of 0.827.

An Ensemble Stock Selection Model | XGBoost, SVM, Random Forest, Soft Voting — 03/2022 - 06/2022

- Conducted data cleaning & exploratory data analysis on 3GB+ stock factor data.
- Labeled stocks and performed feature engineering, selecting important features using IC value.
- Built XGBoost & Random Forest & SVM models, ensembled the 3 models using Soft Voting to predict stock performance, achieving 0.86 ROC-AUC, and successfully selecting top-performing stocks.