ERQIAN (ELSIE) WANG

Jersey City, NJ 07302 | wangeq.elsie@gmail.com | (217) 305-2199 | www.linkedin.com/in/erqian-wang/

EDUCATION

New York University ***Sep. 2022–May 2024***

Master of Science in Data Science

University of Illinois at Urbana-Champaign (UIUC) ***Aug. 2018–May 2022***

Bachelor of Science in Applied Mathematics and Statistics (Double Major) | GPA: 3.99/4.0 (Summa Cum Laude)

Relevant Courses: Statistical Learning, Data Machines and Python, Statistical Modeling, Methods of Applied Statistics, Stochastic Processes, Statistics and Probability, Abstract Linear Algebra

SKILLS

Programming and Tools: Python, R, MySQL, Tableau

Statistics: Linear/Logistic Regression, Ridge/Lasso Regression, Hypothesis Testing, ANOVA

Machine Learning: Decision Trees, Random Forest, XGBoost, k-means, k-nearest neighbors

WORK EXPERIENCE

Data Scientist Intern at PricewaterhouseCoopers (PwC) ***May 2021–Aug. 2021***

Customer Churn Prediction | *Python (pandas, NumPy, matplotlib, seaborn, scikit-learn), SQL*

- Worked cross-functionally with data scientists and product analysts to decrease user churn rate for an e-commerce client using Python.
- Performed data processing (multicollinearity removal, log transformations, standardization, and one-hot encoding) for churn prediction.
- Built and compared Logistic Regression and tree-based models (Decision Trees and Random Forest); tuned hyperparameters for Random Forest; achieved 95% in AUC, 71% in recall, and 94% in precision.
- Evaluated feature importances and assisted in developing a dashboard using Tableau with data automation pipeline ETL for the client to monitor key success metrics on a daily basis, such as DAU, repurchase rate, and customer lifetime value.
- Delivered modeling insights and strategic proposals on churn prevention and promotion to the client.

Data Scientist Intern at Kunlun Health Insurance Company ***May 2019–Aug. 2019***

Health Insurance Fraud Detection | *Python (pandas, NumPy, imbalanced-learn, scikit-learn)*

- Preprocessed data by handling missing values, one-hot encoding, and standardization using Python.
- Utilized SMOTE method to mitigate imbalance in the data by synthesizing new samples for the minority class.
- Constructed XGBoost to detect insurance fraud; increased the recall to 78% (baseline 52%) through oversampling and adjusting sample weights.
- Augmented fraud prevention procedures and projected to reduce losses due to fraud by $3 million annually.

SELECTED PROJECTS

Customer Segmentation with RFM and Clustering | *Python (pandas, NumPy, seaborn, scikit-learn)*

- Aggregated ~500k transaction records into 4k rows for each customer based on the RFM framework.
- Prepared data by standardizing features and removing missing values, duplicates, and outliers.
- Constructed k-means clustering and chose the optimal k value using the elbow method, grouping customers into 4 clusters by their transaction patterns.
- Interpreted clustering results and prioritized customer segmentations for future marketing use.

Email Marketing Effectiveness with A/B Testing | *Python (pandas, NumPy, Plotly, statsmodels)*

- Merged 4 tabular datasets (~1GB) by identifying entity relationships between email campaign datasets deployed to 480k users.
- Defined and computed metrics like email open rate, account linking rate, funding rate, and friction.
- Conducted global hypothesis testing and multiple testing across 24 groups of customers based on their engagement levels; applied Bonferroni correction to alleviate the multiple comparisons problem.
- Visualized conversion funnels to demonstrate 4 fundamental steps in the user journey that lead to funding.
- Made email campaign suggestions based on experimental results.