

Shaohan Chen

917-215-8974 | sc5154@cumc.columbia.edu | New York | www.linkedin.com/in/shaohanchen-orange

EDUCATION

Columbia University New York, United States	Sept. 2022 – May 2024
M.S. in Biostatistics	
Relevant Coursework: Data Science, Statistical Learning and Data Mining, Latent Variable and Structural Equation Modelling, Graphical Models for Complex Data, Bayesian Analysis and Adaptive Designs, Analysis of Longitudinal Data, Relational Databases, Advanced Statistical Computing	
Fudan University Shanghai, China	Sept. 2017 – June 2022
B.S. in Data Science	
Relevant Coursework: Statistical Machine Learning, Object-oriented Programming, Data Structure, Database and Implementation, Data Visualization, Data Mining, Advanced Algebra, Probability, Computational Statistics, Numerical Algorithms, Artificial Intelligence, Mathematical Modelling	

TECHNICAL SKILLS

Programming Skills: Python, R, SQL, Java, Html, CSS, JavaScript, LaTeX
Statistical Modeling: Regressions, Decision Trees, Random Forests, K-Means, KNN, Support Vector Machine, Boosting, Cross – Validation, K – Means, PCA
Data Analysis Tools: A/B Testing, ETL, R studio, Tableau, Excel, Git, Spark

WORK EXPERIENCE

eBay China Analytics Center Shanghai, China	June 2021 – Jan. 2022
Data Science Intern	
<ul style="list-style-type: none">Filtered million class transactional data and computed key business metrics (SQL).Performed visitor funnel analysis based on key metrics such as RUV and BBOWA.Established automated daily and weekly refreshed Excel & Tableau dashboards to track GMV performance of coupons from Afterpay, and upgraded the manual data visualization workflow and save 70% time for building the dashboard.Conducted A/B Testing experiment to simulate monthly GMB performance of eBay Plus Membership, revealed 5+ potentially fruitful aspects to acquire 15% more clients to join plus membership and build closer customer relationship, which provided insights for making eBay's future plus membership plans.Applied K-Means to create buyer segment, adopted regression models (Linear Regression, Lasso, Ridge Regression, GLMNET, R) to predict gift cards performance in different buyer segments, evaluated models based on RMSE and Adjusted R-Squared, delivered model with Adjusted R-Squared > 0.66.Actively involved in projects that adopted uplift models such as T-learner and X-learner (R, Python) for evaluation of seller & buyer coupon uplift benefits.	
Graviti AI Shanghai, China	June 2020 – Sept. 2020
Data Engineering Intern	
<ul style="list-style-type: none">Performed visual analysis reviews on 15+ main autonomous driving datasets (Python) containing 100K image annotation data in 10+ application scenarios, coordinated with algorithm team and made plans to manage data in data warehouse.Developed the SDK of TensorBay and Open Data Platform (Python), engineered data loading functionality for 20K+ image data at back-end database, and enabled image annotation visualization at the front-end website for 10+ published datasets.	
Graph Computing Lab at Fudan University Shanghai, China	Oct. 2019 – May 2022
Research Assistant	
<ul style="list-style-type: none">Researched interactive graph search algorithm (IGS) in graph databases which performed graph search under the guidance of human knowledge, concluded deficiencies of the baseline algorithm studies on the basis of summing up its innovations.Devised knowledge-based interactive graph search (KIGS) algorithm based on entity relation knowledge and weighted-binary search, tested my upgraded algorithm (Python) on multiple datasets and reduced cost of the baseline method by 20%.Published research paper on China Conference on Knowledge Graph and Semantic Computing 2022 (CCKS 2022), gave oral presentation at conference.	
SELECTED PROJECTS	
Credit Card Default Prediction and Analyzing	June 2020 – Sept. 2020
<ul style="list-style-type: none">Explored credit default dataset and conducted analysis of covariance of variables, oversampled the minority class (credit default) using SMOTE algorithm.Implemented classification models (Logistic Regression, Random Forest, SVM, Neural Network, R, Caret, NNet) to predict bank clients' credit default, selected and delivered model with AUC > 0.75, surfaced key insights on clients' transaction behavior patterns and likelihood of credit defaults occurrence.Adopted ensemble learning methods (Boosting, R, Xgboost, SuperLearner) to improve performance, fitted model with AUC > 0.82 and accuracy > 0.84.	
Interactive Movie SQL Database Development	Nov. 2019 – Jan. 2020
<ul style="list-style-type: none">Used the Web Crawler (Python, in legal ways) to extract, collect and organize movie & actor/actress data from 20+ websites on Internet.Established Microsoft SQL Server database, implemented SQL codes to import, clean and tidy extracted raw unstructured movie's data.Engineered functionality for back-end interactive movie data manipulation (such as Create, Delete, Read, Update) using Python and SQL.Crafted front-end interactive interface using PYQT5, built data search, modifying and recommendation functionality at front-end (Pyodbc).	
GOMOKU Chess Artificial Intelligence Player	Nov. 2020 – Jan. 2021
<ul style="list-style-type: none">Built and optimized mathematical modeling towards GOMOKU chessboard game and player placement, investigated some relevant methods.Designed AI chess player by integrating Alpha-Beta Pruning, Monte Carlo, and KMP (Python) algorithm, and beat 85% of other AI players.	
China Air Pollution Data Visualization System	Apr. 2021 – June 2021
<ul style="list-style-type: none">Explored and tidied unstructured China geographical air pollution data, cooperatively designed the interactive architecture of visual analysis system.Constructed visualization system using Html, CSS, JavaScript, and d3.js., analyzed distribution for degree of pollution on the geographic locations.	

PUBLICATION

Shaohan Chen, Weiguo Zheng. *Knowledge-Based Interactive Graph Search*. China Conference on Knowledge Graph and Semantic Computing 2022.