Tengyu Song

st3nv st3nv.github.io tengyu-song ts3464@columbia.edu New York, NY

Key Skills

Data Science
Python, Pandas, Scikit-learn, OpenCV, PyTorch, TensorFlow, Tableau, Machine Learning

Statistics
R, SAS, SPSS, Linear regression, A/B testing, Bayesian analysis, Causal inference

Other
SQL, Git, HTML/CSS, JavaScript, MATLAB, VB.NET, Microsoft Office, Googling

Education

Columbia University, M.A. Statistics, **GPA**: 4.26/4.33 09/2022 – 12/2023
Shanghai University of Finance and Economics, B.Sc. Statistics, **GPA**: 3.80/4.00 09/2018 – 06/2022

Work Experience

Data Analyst, *EGSC Columbia, Part-time* 03/2023 – 12/2023

- Revamped Python data processing codebase for annual QoL survey using Python, Pandas, and NumPy, enabling data processing, analysis, and visualization alongside resolving over 20 existing bugs including data type inconsistencies.
- Created dynamic visualization pipeline by integrating Python (pandas, matplotlib) with Tableau for data transfer via CSV/Excel files, enabling easy adaptation for future datasets and optimizing Python scripts for faster data extraction and transformation.
- Crafted and delivered captivating PowerPoint presentations that highlighted current student needs, leading to a widely attended townhall and a publicly acclaimed report.

Data Engineer Intern, *JD.com, Internship* 09/2021 – 11/2021

- Managed daily data extraction tasks using SQL and designed a comprehensive visualization dashboard by employing Tableau and Power BI that illuminated essential performance metrics in form of charts, graphs, and tables, enhancing operational efficiency for trend analysis.
- Designed and implemented machine learning-driven coupon grading system using machine learning model's GBDT/Random Forest for enhanced decision making, resulting in 10% surge in company's 7-day Gross Merchandise Volume (GMV) .
- Led anomaly detection project on R&D process flow data related to process times, sequence patterns, or deviations from expected behaviors and leveraged Isolation Forest, One-Class SVM, Local Outlier Factor (LOF) , and k-nearest neighbors (KNN) , achieving an 80% detection accuracy.

Data Analyst Intern, *Yum China Holdings, Internship* 04/2021 – 08/2021

- Authored and executed data queries using SQL with Hive and Impala including predictive modeling using window functions, time-series analysis, and cohort analysis to derive actionable insights for decision-making, accumulating over 1000 lines of SQL code.
- Investigated user behavior hypotheses based on qualitative insights and stakeholder inputs accessed via web logs, databases, or user interaction logs using rule mining, sequence analysis, or regression analysis, delivering crucial insights for strategic decisions.
  Led mini project Cross Analysis of User Acquisition and Retention, identifying critical factors in customer engagement (user acquisition sources, conversion rates, customer churn rates, and retention rates) , improving company's SMS marketing efficiency by 20%.

Selected Projects / Research

Dynamic Pricing for MTA Subway System, *Python, Scikit-learn, Statistical Analysis*

- Introduced a dynamic hybrid pricing model consisting of Random Forest and ARIMA model to solve the issue of over-crowdedness and increase fare revenue. Also addressed the issue of price discrepancy in nearby stations by implementing spectral clustering on geographical and ridership data.

Automatic Defect Detection of PV Cell Panels, *Python, OpenCV, PyTorch, Computer Vision*

- Performed multi-category defect detection on photographed images of PV cell appearance using finetuned Mask-RCNN model.
- Devised an effective image cropping algorithm using OpenCV to reduce detection difficulty.

ChatGPT vs Human on Coding Problems, *Python, PyTtorch, Huggingface, LLMs*

- Used API to collect ChatGPT's answers on multiple coding problem, assessing the model's accuracy on different tasks in the data science domain.
- Developed RoBERTa-based NLP model to distinguish between human-authored and GPT-generated code snippets.

SUFE Rating Desktop Version, *Database, SQL, VB.NET*

- Led a team of 15 that created the desktop version of most influential professor rating platform on campus, with up to 20,000+ users and 15,000+ highest number of visits in a single day.

More Projects / Research

Bagging Enhanced Sparse Recovery Algorithms, *Python, Numpy*

- Optimized signal recovery performance by applying bagging techniques to refine Orthogonal Matching Pursuit and Matching Pursuit algorithms.
- Executed rigorous experiments on simulated datasets. Evaluated performance of different bagging strategies and tuned hyperparameters to identify the optimal configuration.

Differential Privacy under Robust M-estimators

- Formulated a concentration bound for sensitivity curve of M-estimators based on robust statistics theory.
- Further simplified the noise tuning process for robust estimators in (ε, δ)-differential privacy framework by building connection with the smoothed local sensitivity of datasets.

Microblog Rumor Diffusion and Debunking Dynamics, *Python, Beautiful Soup, Pandas*

- Carried out data collection process. Crawled over 10,000 sets of rumor-related content from Weibo using Python and Beautiful-Soup.
- Using SIIR model to simulate counter-propagation rumor spreading dynamics, employing Random Forest to quantify the influence of various determinants in rumor mitigation.

Extracurricular / Teaching Experience

Teaching Assistant, *Introduction to Statistic* 01/2023 – 05/2023

- Hosted bi-weekly office hours, providing tailored assistance for students' homework and final project, boosted students' performance in the final exam by 5%.
- Collaborated with the instructor to devise comprehensive solutions for students and detailed rubrics for grader, ensuring fast and transparent grading processes.
- Developed lab materials on R data analysis and conducted monthly lab sessions to 80 students, highly praised by students and faculty.

Mentor Leader, *Summer Training for Aspiring Researchers (STAR) Program* 05/2023 – 07/2023

- Led the mentor group to provide academic support for undergraduate mentees. Designed and delivered innovative math/statistics refresher and problem sets.
- Organized social engagement events, such as board game afternoon and chess breaks, enhancing connections between mentors and mentees.

Selected Honors / Awards

Chair's List of Academic Achievement	2023
JSM Award	2023
First Place of Columbia Data Science Hackathon	2023
People's Scholarship	2019