Baipeng Gong

New York, NY, 10011 | (646)877-8364 | bg1622@nyu.edu | LinkedIn Profile

EDUCATION

New York University September 2021 – May 2023
Master of Science in Data Science **GPA: 4.0/4.0**

- **Anticipated Coursework:** Introduction to Data Science (A/B Testing), Machine Learning, Big Data, Natural Language Processing (NLP), Deep Learning, Probability and Statistics, Optimization and Linear Algebra, Time Series Analysis

New York University September 2016 – May 2020
Bachelor of Science in Data Science, Finance **GPA: 3.71/4.0**

- **Honors:** Cum Laude, University Honors Scholar, Founders Day Award, Dean's List
- **Coursework:** Data Structures, Database Systems, Information Visualization, Numerical Analysis, Econometrics

SKILLS

Programming Languages: SQL, Python (NumPy, SciPy, Pandas, Matplotlib, Scikit-learn, PyTorch, TensorFlow), MATLAB
Tools: Microsoft Office, Tableau, Git, Hadoop, Spark, HPC, Stata, LaTeX, iMovie

PROFESSIONAL EXPERIENCE

Morgan Stanley **New York, NY**
ESG Data and Analytics Summer Analyst June 2022 – August 2022

- Tried multiple data manipulation tricks, Python NLP libraries (including spaCy and NLTK), and evaluation metrics to match the company names between REITs dataset and Green Building Certification datasets **(over 100K records)**
- Built a language model using CountVectorizer for stop word detection and removal, N-gram for tokenization, and TF-IDF for embedding to find the best match between two lists of company names by comparing cosine similarity
- Established **a new scheme** that applied fuzzy string matching to compare text similarity by fuzzy ratios based on the Levenshtein Distance algorithm, **simplified the model and made it easier to understand for non-technical audiences**
- Leveraged the waterfall methodology to combine the above two schemes, **increased the number of matches by 80% and the matching accuracy to around 90%**
- Proposed using set theory and numerical analysis thinking to adjust the filtering threshold, **improved the flexibility and reliability of filtering than heuristic threshold selection**
- Analyzed the differences between the matching results of 3 Green Building Certification datasets, **obtained new findings and presented them to the Vice President, laid the foundation for the follow-up factor investing**
- Visualized Green Buildings on Google Map and drew choropleth maps using Google APIs and Python libraries such as Bokeh and GeoPandas

Analysys International **Beijing, China**
Data Analytics Intern June 2020 – December 2020

- Quantified business for **10+ clients** from various industries, determined performance metrics for each client
- Deployed web analytics tags for data collection, participated in building pipelines using SQL and Python, **sped up data processing time by 25%**
- Delivered data reports and dashboards with a special focus on Event Trends, Conversion Funnel, and Retention Analytics
- Provided clients with suggestions for improvement, monitored A/B Testing, **achieved significant growth in Daily Active Users, Page Views, Clickthrough Rate, Gross Merchandise Value, etc.** for different clients
- Wrote articles summarizing the optimization strategies for webpages/apps, posted them on company's WeChat Official Account, received **3000+ views** for some articles

RESEARCH & SELECTED PROJECTS

Improving Numerical Reasoning Skills for Financial QA February 2022 – May 2022

- Performed intermediate training with MathQA dataset to improve the automations of financial report analysis, **achieved accuracy 10% higher than general crowd performance**
- Utilized GenBERT and TASE-BERT encoders as drop-in replacement for BERT-base and RoBERTa-large models in the Financial QA architecture, **improved the accuracy by 1.5% than the baseline framework**
- Conducted qualitative analysis and investigated the errors, found that our model performs better for table-only questions and questions that require less than 3 operations

Evaluating COVID-19 Vaccine Hesitancy among People September 2021 – December 2021

- Cleaned **1.5M+ user data** with **300+ features**, performed feature engineering with forward selection and one-hot encoding
- Constructed Logistic Regression model to identify the people who are most likely to be hesitant to get the COVID-19 vaccine, reached a **0.83 AUC score**
- Applied PCA and K-Means to cluster users into 6 groups, analyzed potential reasons people are not receiving vaccination