YUXUAN FU

(646) 462-0849 | yf2592@columbia.edu | Linkedin

EDUCATION

Columbia University New York, NY

M.A. in *Statistics* Sep 2021 – Dec 2022

- Courses: Statistical Computing & Data Science, database design, deep learning, Algorithm Analysis, Cloud Computing & Big Data, NLP
- Leadership: Columbia University Chinese Students and Scholars Association – Media and Public Relations member

East China Normal University Shanghai, China

B.S. in *Statistics* Sep 2017 – Jul 2021

- University Outstanding Scholar-2nd prize; Supervisor of department's Student Office, Practice & Innovation Center
- Courses: Sampling Survey, Machine Learning, Bayesian, Stochastic, Experimental Design, Nonparametric, C++, Business Analytics

University of California, Berkeley - *Exchange Student* Berkeley, CA

Courses: Statistical Computing (graduate level), Time Series, Linear Analytics and Regression. Sep 2019 - Jan 2020

PROFESSIONAL EXPERIENCE

Children and Screens: Institute of Digital Media and Child Development New York, NY

Research Assistant Volunteering Jan 2022 - Present

- Review latest academic literatures for scholar database and weekly digest presentation, minute webinar and write YouTube summaries.
- Analyze and visualize attendance/outreach/feedback webinar data with Tableau and Demo to offer outreach improvement advice.

Roche Diagnostic Shanghai, China

Data Science Intern - Product Development Global Clinical Execution Team (PDG) Jun 2020 – Mar 2021

- Devised a data visualization platform with clinical management team to promote over 10 corporate-scale R&D tasks.
- Transmitted data with R and python, including reports crawling from Almac and Rave database, R preprocessing, Chrome GAuth tokens' authentication and Google Data Studio interactive template design.
- Specified 7 - 10 paged dashboard templates respectively for six medical experiments by daily communicating with CRAs.
- Probed advanced NLP tasks with CDM to mine the similarity of drug description text, using LDA and deep learning algorithm.

Accenture Shanghai, China

Project Assistant – IT Consulting Department Jan 2020 – Mar 2020

- Provided customer with IT support, cloud service and cloud office for an energy development project.
- Completed over 40 documents' content organization, formatting, partial proofreading, translation and data sorting.
- Integrated a 2GB team scheme based on customer inquiries and fulfilled the customer delivery task with manager and partners.

RESEARCH PROJECTS

Multi-model human face expression video recognition (Python) Oct 2021- Dec 2021

- Operated human facial expression recognition system for video data, applying retina net for face detection, face alignments.
- Literature review for VGG-n in combination with LSTM architecture to reduce hyperparameter numbers and raise accuracy.
- Gained fusion score with C3D, VGG16_LSTM and achieved prediction results with C3D network on Google Cloud Platform.

Anomaly Detection of Civil Aviation Flights based on QAR data (R) Jan 2021 – Jun 2021

- Preprocessed 49 flights QAR dataset within 7944 time-varying variables through parallel computation, considering time constraints, strong interference, operational meanings and rough-set's encoding patterns.
- Visualized 644 variables' groupings to respectively 35 and 45 groups with multi-time series method: autocorrelation-function-based Fuzzy C-means Clustering combining SBD distance and K-means Clustering combining DTW distance.
- Implemented KPCA for representative feature extraction and detected 7-10 potential abnormal flights with single variable's functional boxplots and feature-vector based DBSCAN clustering.

Infants' Sleep Quality, Neurobehavioral Development, Language & Social Interaction Research (R) Jan 2020 – Apr 2021

- Collaborated with hospital statistics of more than 200,000 standardized questionnaires across nation, measured the effects of gestational age at birth on infants' sleep quality, neurobehavioral development, and language and social interaction.
- Experimental design and linear regression, logistic regression, nonparametric and multilevel statistical analysis modeling with continuous and discrete dependent varies, independent varies and covariates. Paper in submission for publication.

Prediction of Bitcoin Price using Machine Learning and Sentiment Analysis (Python) Sep 2020 – Nov 2020

- Extracted sentiment from 4.9M Twitter & Reddit comments with VADER, Text-Blob and LM dict. as prediction's proxy.
- Integrated time by weighted-moving averaging sentiment polarity & subjectivity, bitcoin close price, and platforms' popularity index.
- Extracted 88% explanation of variability feature with LightGBM to simplified DL network's Input.
- Forecasted data trend with LSTM, LSTM_CNN, GRU and identified GRU as better with 62% lower test RMSE than other two.

User-Generated Content (UGC) Information Accuracy Evaluation and Influencing Factor Analysis (Python) Mar 2019 – Apr 2020

- Directed the independent subject on UGC accuracy evaluation based on Baidu, Insight-China Pharma database.
- Procured over 1000 opensource medical prescriptions' text quickly by utilizing *urllib, BeautifulSoup* for Baidu webpage and *ajax, selenium, requests* for database, overcoming asynchronous login and JavaScript obfuscator issues.
- Implemented LDA combining perplexity, KL divergence model in self-written functions and parallel preprocessed 6 GB unstructured text data, including text segmentation, word tagging and keyword extraction.
- Visualized text comparison of html output with *difflib* through python and terminal commands.

TECHNIQUE SKILLS

- **Languages:** R, Python, SQL, PySpark, SAS, C++, Ubuntu, Bash Shell, Git, JavaScript, Latex
- **Big Data Toolkits:** MapReduce, Snowflake, GCP, AWS, Azure | **ML Toolkits:** Scikit-Learn, TensorFlow, MXNet
- **Data Visualization:** Tableau, R Shiny, Google Data Studio, Power BI | **Methodology:** A/B testing, experiment design, data mining