Yun Wang

Tel: +1 (706)-461-7882⋄ Email: Yun.Wang1@uga.edu⋄Homepage: https://yunnw.github.io/

EDUCATION

University of Georgia — **Georgia, U.S.**
Doctor of Philosophy in Computer Science — **Aug 2024 – Present**
- Advisor: Prof. Ninghao Liu
- Research Interests: Large Language Models, Foundation Models, Trustworthy AI

University of Leeds — **Leeds, U.K.**
Master of Science in Data Science and Analytics — **Sept 2019 – Nov 2020**
- GPA: 3.8/4.0 (Pass with Distinction)
- Modules: Machine Learning, Data Mining and Text Analytics, Linear Regression and Robustness, Generalized Linear Models, Statistical Computing, Big Data Systems, Data Science, Knowledge Representation and Reasoning

Jiaxing University — **Zhejiang, China**
Bachelor of Science in Applied Statistics — **Sept 2014 – Jun 2018**
- GPA: 3.6/4.0
- Modules: Multivariate Statistical Analysis, Bayesian Statistics, Database System, Applied Stochastic Processes, Statistical Data Analysis, Probability Theory & Mathematical Statistics, Statistical Forecasts and Decision Making

SKILLS

- Front-end Development: HTML/CSS, JavaScript, React
- Backend Development: Node.js, Express.js, Django, Ruby on Rails
- Data Analysis and Machine Learning: Python, Pandas, NumPy, Matplotlib, Seaborn, TensorFlow, PyTorch, Scikit-Learn
- Database Management: MySQL, MongoDB, PostgreSQL
- Cloud Services: AWS, Azure, Google Cloud
- Version Control and Code Collaboration: Git, GitHub, GitLab, Slack, Microsoft Teams
- Operating Systems: Linux, macOS, Windows
- DevOps Tools: Ansible, Terraform

REPRESENTATIVE RESEARCH EXPERIENCE

Master Thesis: Apply Machine Learning to Identify Shallow Cumulus Cloud Types
Supervised by Professor Douglas Parker — **Sept 2020 – Nov 2020**
- Focused on classifying shallow cumulus clouds, the dissertation advances climate prediction by addressing their impact on Earth's radiation balance.
- Employed innovative data processing techniques and convolutional neural networks, with a unique use of run-length encoding for data representation and augmentation.
- Achieved an 89% accuracy in classifying various cloud types, highlighting the model's effectiveness and challenges.
- Provided insights into the model's development, potential applications, and limitations, paving the way for climate enhancements.
- Significantly contributes to environmental and climate science, showcasing machine learning's potential in understanding atmospheric processes and the global climate system.

REPRESENTATIVE PROFESSIONAL EXPERIENCE

Beyondsoft Shanghai Co., Ltd — **Shanghai, China**
Data Scientist / Machine Learning Engineer — **Dec 2021 – Jul 2024**
- Led extensive data collection from diverse sources, both structured and unstructured; Managed end-to-end ML and software projects using agile methods.
- Applied advanced algorithms to address complex business challenges, fine-tuning machine learning models for predictive analytics, recommendations, and NLP.
- Engineered informative features from raw data, leveraging domain knowledge to enhance model performance.
- Collaborated effectively with cross-functional teams, translating complex concepts for non-technical stakeholders; Stayed updated with the latest ML advancements.
- Proficiently deployed models in production environments, including RESTful APIs using Flask and web application development.

REPRESENTATIVE PROJECTS

Comprehensive Research and Enhancement of Language Models for Natural Language Processing
- Conducted in-depth research on Transformer-based language models, including BERT and GPT, harnessing their pre-training capabilities for complex language understanding.
- Employed data augmentation techniques to handle unstructured data, leveraging NLTK and spaCy for tasks like Part-of-Speech Tagging and Named Entity Recognition.
- Utilized high-dimensional word embeddings and contextual embeddings for deep word meaning capture.

- Explored various models, from logistic regression to ensemble methods like XGBoost, and fine-tuned hyperparameters with Bayesian Optimization.
- Addressed class imbalance with techniques like SMOTE and ADASYN for better model generalization.
- Implemented model pruning and utilized TPUs and GPUs for efficient training and improved system performance.

Advancing Customer Engagement with GPT-4-Based Chatbot, Partner Center Copilot

- Applied advanced text clustering techniques, including LDA and BERTopic, for topic modeling on extensive customer case and ticket data, identifying themes and clustering related documents.
- Fine-tuned the BERTopic model for nuanced context differences and improved accuracy.
- Enhanced model quality iteratively with NLP techniques like part-of-speech tagging and dependency parsing.
- Integrated NLU to align model labels with internal documents and improve chatbot responses.
- Carefully crafted prompts for accurate and company-aligned chatbot responses, establishing a feedback loop for ongoing model refinement based on user interactions.

Enhancing Customer Support Through Data Analysis and Knowledge Base Creation

- Analyzed customer service cases and ticket data for insights into issues with Microsoft Partner Center. Established a knowledge base framework with answers to common questions, service processes, and product information to support GPT-driven assistance.
- Collaborated with subject matter experts to structure and populate the knowledge base using natural language processing techniques.
- Extracted patterns and question-answer structures from user inquiries, establishing an efficient knowledge management system.
- Designed effective prompts to connect the chatbot to the OpenAI API for problem-solving using the knowledge base information.

Impact Analysis of the Service Product ASfP

In-Depth Technical Analysis and Causal Inference:

- Conducted ASfP (Advanced Support for Partners) service impact analysis, emphasizing technical intricacies.
- Utilized Propensity Score Matching (PSM) to select a well-matched control group from clients who hadn't used ASfP.
- Employed Difference-in-differences (DiD) to quantify the uplift effect of ASfP service using specialized software packages.

Time Series Analysis and Predictive Modeling:

- Conducted advanced time series analysis utilizing the LSTM model with TensorFlow and PyTorch.
- Developed a customized LSTM model in Python to predict outcomes for clients without ASfP(Advanced Support for Partners) service, showcasing its innovative predictive capabilities through performance comparison with actual results.

Statistical Validation, Hypothesis Testing, and Data Visualization

- Usedtatistical analysis, including specialized software like R and SAS, to demonstrate the positive impact of the ASfP service; Employed advanced hypothesis testing, such as bootstrapping and Monte Carlo simulations, to confirm statistical significance.
- Crafted an interactive and visually engaging Power BI dashboard, presenting an overview of the analysis results and descriptive insights, highlighting the technical complexity of the project and providing actionable insights for the ASfP team and stakeholders.

Advanced Data Handling and Automation Techniques in Power BI Integration

- Utilized REST API in Power Automate for data access and manipulation, enhancing code maintainability.
- Employed OfficeScript and VBA scripts for data transformation and scalability.
- Implemented a custom data refresh mechanism using JavaScript for dynamic updates.
- Utilized asynchronous programming, caching, and WebSocket technology for improved performance and real-time data updates.
- Developed an event listener for Power BI dataset event streams, gaining expertise in WebSocket protocols and event handling.

Innovative Approaches to Enhancing LSTM for Stock Market Prediction

- Explored statistical techniques for stock market predictions, focusing on LSTM's efficiency and generalizability.
- Introduced a single-layer neural network and discrete continuous data blocks to enhance LSTM's trend learning.
- Addressed LSTM training challenges by reducing memory usage and improving prediction performance.
- Optimized the fully connected neural network for complex relationship capture in time series data.
- Utilized discrete continuous data blocks for improved long-term trend forecasting in the stock market.

AWARDS

- ✧ Beyondsoft Excellent Employee, Dec 2022
- ✧ Top 20 China University SAS Data Analysis Contest, National Final Dec 2017
- ✧ First Place China University SAS Data Analysis Contest, Zhejiang Division Nov 2017
- ✧ First Prize 6th Zhejiang Provincial University Students' Statistical Survey Scheme Design Contest Nov 2017