Jiawei Wang

EDUCATION

Master of Science in Business Statistics 08/2022 – 12/2023
Washington University in St Louis St. Louis, MO
GPA: 3.9/4.0, Concentration: **Data Analytics, Machine Learning**
Beta Gamma Sigma Award

Bachelor of Arts in Mathematics 08/2018 – 05/2022
University of Colorado Boulder Boulder, CO
GPA: 3.3/4.0, Minor in **Economics and Business**, Concentration: **Statistics**

RESEARCH EXPERIENCES

Improving Bunge's Rail Car Fleet Sizing Model **The Boeing Center**
Supervisor: Dr. Panos Kouvelis 2023/01 – 2023/05

- Introduced and validated a new data collection workflow that accelerated data gathering while ensuring utmost accuracy, resulting in a remarkable 15% improvement in data collection efficiency
- Redesigned and added new constraints to the existing model, leading to a 17% increase in fleet utilization, reaching up to 91%, and adeptly adapting to fluctuations in commodity demand and rolling stock supply
- Established a backtest framework and trained the LSTM neural network to continually improve the accuracy and precision of the model, achieving 87% accuracy in fleet forecasting
- Awarded the honor of Best Innovation by the Boeing Center

An Interpretable Artificial Intelligence Tool for Mental Health Screening **Washington University in St. Louis**
Supervisor: Dr. Salih Tutun 07/2022 – 03/2023

- Collected and cleaned data feedbacked by SCL-90-R, created a feature space, and projected high-dimensional features on a two-dimensional plane using **t-SNE**
- Performed multiple classification on two-dimensional images with **CNN**, identifying 10 mental disorders, and trained a CNN model with best performance
- Generated interpretative visualization interfaces for each prediction using interpretation algorithms such as **SHAP** and **LIME**, highlighting key features and contributions

Application of graph networks to mental health data **Washington University in St. Louis**
Supervisor: Dr. Salih Tutun 07/2022 – 12/2022

- Preprocessed the survey data by converting the respondents' binary answers into features on nodes and visualized the relationships between questions with directed edges in a graph
- Applied network analysis methods to calculate various centrality indices such as degree, betweenness, and closeness of each node based on the graph topology and defined them as topological features of nodes
- Constructed a multi-layer graph convolutional network model, including graph convolutional layer, fully connected layer, etc., to study the complex interaction between problems
- Trained the model using a mental illness dataset containing annotations of answers, achieving 85% accuracy on the test dataset
- Visualized and conducted feature importance analysis to optimize the model's interpretability

Solar Flare Frequency Distribution Analysis **UCB Atmospheric and Space Lab**
Supervisors: Dr. Heather Lewandowski and Dr. Colin West 08/2021 – 12/2021

- Conducted in-depth **statistical analysis** on a large dataset of 8M+ solar flare observations from 1980-2015 to uncover distribution patterns and intricacies
- Developed a statistical model using **Gaussian likelihood functions** and Markov Chain Monte Carlo (**MCMC**) simulation for precise parameter estimation
- Published research paper on findings related to nanoflares and coronal heating mechanisms and validated the superiority of Gaussian/MCMC approach over baseline methods

ACADEMIC PROJECTS

Designing an Optimized Regional Distribution Center Network for Dartboard, Inc. **Washington University in St. Louis**
Supervisor: Dr. Amr Farahat 08/2023 – Present

- Collected and analyzed weekly sales data with over 3 million records from 765 counties during the past 3 years using R
- Constructed a regression model integrating time series analysis, seasonal adjustment, and logarithmic conversion to forecast weekly sales for each county, with a MAPE value of 12%

Jiawei Wang

- Evaluated 17 alternative distribution locations, designed an integer programming model using Python to minimize construction and operating costs, and determined on building five new distribution centers, increasing the capacity by more than 20%
- Adopted GIS to calculate the transportation distance and cost between counties and distribution centers, realizing total cost savings of more than 5%

Statistical Modeling and Analysis of Traffic Accident Severity **University of Colorado Boulder**
Supervisor: Dr. Joseph Timmer 2022/01 – 2022/06

- Analyzed a dataset of 2.8M U.S. car accidents using **PySpark** to identify key factors contributing to fatal crashes and performed distributed data cleaning, exploratory analysis, and feature engineering
- Developed machine learning models to predict crash severity, including **random forest**, **XGBoost**, **SVM** and **logistic regression**, and tuned hyperparameters using **grid search** and cross validation to optimize accuracy, recall and F1-score
- Created interactive **Tableau dashboards** to visualize key trends and patterns in fatal accidents across regions and demographics, highlighting priority areas for safety interventions

WORK EXPERIENCE

Data Scientist, Schnucks **Saint Louis, MO | 09/2023 – Present**

- Collect and clean customer buying behavior data from questionnaires, visualize high-dimensional data using technologies such as t-SNE, identify customer buying patterns and segment the customer base
- Designed and trained a dedicated convolutional neural network (CNN) model to predict customer loyalty based on customer characteristics, achieving 81% accuracy
- Integrate deep learning frameworks such as Vision Transformers, **ResNet-101** and **ResNeXt** using **PyTorch** framework to build high-performance hybrid models, improving the model's accuracy to 87%
- Interpret model predictions with algorithms such as SHAP and determine the relative importance of each feature to the prediction results, increasing model transparency and helping managers better understand and use the model

Data Scientist Intern, IntelliPro Group **Santa Clara, CA | 05/2023 – 08/2023**

- Engineered **ETL** pipelines using **Python** to extract web data to **AWS S3,** improving data processing efficiency by 40%
- Enhanced and refined **LLM** model at **LangChain** through prompt engineering and fine-tune techniques, leading to more accurate results, a 15% improvement in candidate screening, and an $11k reduction in costs
- Designed a **job recommendation system** using Graph Neural Networks, enhancing user experience, increasing match accuracy with **PyTorch** Geometric, and reducing manual job matching effort by 10 h/w
- Developed a machine learning classification system using **Hugging Face's** Transformers library to automatically categorize academic papers into certain fields

Data Analyst Intern, Bunge Limited **Saint Louis, MO | 01/2023 – 05/2023**

- Built optimization models in Python utilizing **Linear Programming** to optimize $2M rail fleet size and distribution, reducing costs by $250K/year and improving fleet utilization by 10%
- Leveraged **ARIMA and LSTM neural networks** to accurately forecast quarterly customer demand at railyards, achieving a high prediction accuracy rate of less than 8% error (MAPE)
- Streamlined and automated the workflow for demand forecasting, fleet management, and logistics planning using **VBA**, leading to a more efficient decision-making process
- Presented analysis and model results to executives, facilitating their data-driven decision-making

EXTRACURRICULAR ACTIVITIES

Advanced teaching assistant，Prescriptive Analytics, Washington University in St. Louis **10h/w; 10/2023 – 12/2023**

- Graded students' assignments and tutoring during the lab session each week
- Attended office hours each week to solve doubts students encountered in learning

Teaching Assistant, for Machine Learning coursework, Washington University in St. louis **20h/w; 08/2022 – 12/2022**

Learning Assistance, Statistical Inference, the University of Colorado Boulder **20h/w; 08/2021 – 12/2021**

SKILLS

- **Programming**: Python, R, SQL, AWS (Redshift, S3, SageMaker), Hadoop, Microsoft Excel, Tableau, Apache Spark, LaTeX
- **Machine Learning**: Deep-Learning (CNN, RNN, LSTM, VAEs), Regression (Linear, Logistic), Classification (SVM, Naive Bayes, Decision Trees), Clustering (K-Means, Hierarchical Clustering), Dimensionality Reduction (PCA), Feature Engineering, Time-Series (ARIMA, Holt-Winters), Hypothesis Test, A/B Testing, GLM