Binrui Yang

by2361@columbia.edu | 408-966-5168 | Linkedin.com/in/binrui-y

EDUCATION

Columbia University – New York, NY **Sept 2023 – Dec 2024 (expected)**
Master of Arts in Statistics (STEM), Graduate School of Arts and Sciences (GPA: 3.96/4.0)

University of Wisconsin-Madison – Madison, WI **Sept 2019 – May 2023**
Bachelor of Science, Double Major in Statistics and Mathematics (GPA: 3.55/4.0)

PROFESSIONAL EXPERIENCE

Travelers – Hartford, CT June 2024 – Aug 2024
Data Science Intern, Auto/Telematics Team in Personal Insurance Research & Development

- Developed an **XGBoost** classifier to predict IntelliDrive policy-level sufficiency using UPP quote data from over 5 million drivers, enhanced with **Optuna**'s customized objective function, boosting 9.6% accuracy of identifying sufficient cohort.
- Crafted an interactive dashboard using **Matplotlib** and **Plotly** to visualize the driver's program completion, leading to a 15% improvement in pinpointing drop-off points and enhancing intervention strategies.
- Automated an end-to-end model pipeline with **Apache Airflow** and deployed the XGBoost classifier using customized Amazon EC2 and S3 instances, improved the real-time prediction capabilities by 18.4% at Rate Call 1.
- Coordinated with cross-functional teams and presented project progress and findings to R&D and business partners, resulting in the adoption of key model insights into future product strategy.

Eth Tech – Newark, CA June 2022 – Aug 2022
Data Analyst Intern

- Utilized **SQL** and **Tableau** for comprehensive data analysis to provide insights into **marketing funnels and business KPIs** for an online marketplace; improved overall Weekly Active Users (WAU) by 3%.
- Employed **Logistic Regression** and **Random Forest** models on a dataset of 10M+ users to forecast conversion rates, achieving a 91% prediction accuracy as validated by AUC.
- Designed and executed **A/B experiments** on user conversion **funnel flow** and leveraged **SQL** for detailed user journey analysis, resulting in a 5% improvement in user conversion rate and 3% reduction in churn.
- Collaborated with marketing team to segment users into 20 treatment groups and devised personalized marketing campaign strategy; with initial results indicating potential for further increases in WAU and overall user engagement.

Epistemic Analytics – Madison, WI March 2022 – May 2023
Data Analyst Intern

- Developed a data simulation algorithm using **Python** to generate diverse, stakeholder-specific land-use scenarios from log data for a land-use planning simulation game *iPlan*, enabling a deeper understanding of user decisions in a simulated environment.
- Applied Singular Value Decomposition (**SVD)** for feature selection and dimensionality reduction, simplifying high-dimensional data into a two-dimensional metric space for analysis, achieving 92% explained variance.
- Created interactive data visualizations using **Plotly R** to represent and interpret land-use scenarios, enhancing the interpretability of user's problem-solving processes.
- Applied statistical analysis to map stakeholder satisfaction; utilized a measurement model to project new user-generated scenarios, offering actionable insights into learners' decision-making processes for educators.

PROJECTS

2024 Travelers NESS Statathon - Gold Medal Winner May 2024

- Collaborated with a team of four to develop an auto quote customer conversion model using **stacking** with tuned **XGBoost**, **LightGBM**, and **CatBoost**; achieved the highest prediction accuracy (**AUC**) among all participating teams.
- Conducted extensive **data analysis**, **data cleaning**, and **feature engineering** to prepare the dataset for modeling, analyzed and identified key characteristics and trends of policies with high and low conversion rates.
- Provided strategic recommendations on leveraging the model insights for improving policy conversion rates.

Text Analysis on Amazon Reviews April 2023

- Prepared and processed 500K+ review records text data using **Numpy**, **Pandas**, **scikit-learn**, and **NLTK** libraries in **Python**.
- Performed data cleaning and constructed a term-doc incidence matrix and generated word clouds for automated sentiment analysis of user experience and satisfaction ratings within product listings.
- Utilized **N-Gram** and **TF-IDF** for text vectorization; trained **logistic regression** and **random forest**, achieving 96% accuracy as measured by AUC with cross-validation and testing.

RELEVANT COURSEWORK

Advanced Machine Learning, Deep Learning and Neural Networks, Big Data Systems, Bayesian Analysis, Probability Theory, Statistical Inference, Applied Regression Analysis, Natural Language Processing, Design and Analysis of Online Experiments

SKILLS

- **Programming & tools:** Python (Numpy, Pandas, Scikit-Learn, TensorFlow, PyTorch), SQL, Spark, Kafka, Git
- **Modeling:** GLM, Decision Tree, Random Forest, KNN, GBM, SVM, K-Means Clustering, NLP, Neural Networks
- **Data Analysis:** Matplotlib, Seaborn, Shiny R, Plotly R, ggplot2, Tableau, Power BI, A/B Testing & Experimentation