JunJie (J.J.) Zhang Ph.D.

JJ.ZHANG.IN@gmail.com | Phone: (317) 532 7526 | GitHub: JJ-Zhang-DS | LinkedIn: in/zhangj-j/ | US citizen

SUMMARY

- Data Scientist with 6+ years of experience developing and implementing machine learning algorithms to analyze large datasets, solve complex business problems, and deliver data-driven solutions.
- Proficient in predictive modeling, statistical analysis, and GenAI, with a strong ability to collaborate across teams and effectively communicate technical insights to stakeholders.
- Experienced in leading innovative AI-driven solutions that enhance decision-making processes and operational efficiency.

SKILLS

Programming Languages: Python, R, SAS. **Database & Cloud Tools**: SQL, GitHub, Azure, AWS, SageMaker. **AI Techniques**: Supervised and Unsupervised Machine Learning, AutoML, Deep Learning, PyTorch, Scikit-Learn, Forecasting, Time Series, Classification, Regression, GenAI, RAG, Natural Language Processing (NLP). **Statistical Analysis**: Experimental Design, Correlation Analysis, Hypothesis Testing, Power Calculation, Variance Analysis.

EXPERIENCE

Bayer AG Indianapolis (remote), IN

Senior Statistical Data Scientist (contract) 10/2022 - present

Advanced Crop Yield Forecasting Model – Commercial Production

- Collaborated across domains to drive the development of seed production pipelines, generating insights that guided market decision-making.
- Applied various machine learning and statistical techniques, including Best Linear Unbiased Prediction (BLUP), Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Neural Networks, to predict crop yield using historical, parental testing, genomic, phenomic, and environmental datasets.
- Developed an ensemble model that enhanced forecasting accuracy and introduced innovative digital solutions for commercial seed production, resulting in 3,000 acres of land savings annually.

Seed Quality Classification and Emergence Vigor Prediction – Research and Development (R&D)

- Led ‘Lab to Land’ R&D studies using advanced statistical and machine learning approaches to refine seed quality management processes.
- Developed a k-means multi-class classification system to optimize seed quality assessments, enabling evaluation of emergence risks across diverse environments and optimizing seed distribution strategies. This improved marketable seed by 5% annually and boosted customer satisfaction.
- Utilized RF to predict emergence from quality lab tests, achieving an accuracy of 0.85. In-depth analysis revealed the test requirements could be reduced by 50% without compromising accuracy, significantly lowering assessment costs.
- Applied Linear Mixed Models and simulations for experimental design and power analysis, reducing experiment costs by 20% without loss of power.

GenAI Chatbot and Comments Categorization

- Developed a Retrieval-Augmented Generation (RAG)-based chatbot leveraging generative AI to provide internal users with instant, context-aware responses on domain knowledge and locating document sources.
- Designed and fine-tuned the chatbot, integrating internal documents, scientific literature, and genomic datasets for enhanced knowledge accessibility.
- Refined the model based on user interactions, reduced onboarding time for new employees by 60% and automated 80% of common user queries, minimizing reliance on manual documentation and expert consultations.
- Built a BERT engine to categorize operation comments during planting season. Pipelined text extraction, automating text extraction and conducting sentiment analysis to generate risk alerts for potential quality issues.

Indiana University School of Medicine Indianapolis, IN

Assistant Research Professor/ Data Scientist 07/2017 - 03/2022

Genetic Mutation Identification from Literature Review with NLP

- Developed a Natural Language Processing (NLP) pipeline to detect gene and mutation mentions from large-scale biomedical texts, processing over 5,000 scientific articles to identify potential disease-associated mutations.
- Applied sentiment analysis to classify whether mutations were positively or negatively regulated in neural diseases, achieving a 90% classification accuracy and enabling 5x faster of mutation identification for further experimental and clinical research.

Innovative Anti-Hepatitis B Virus (HBV) Drug Discovery:

- Utilized unsupervised machine learning to classify compounds based on structural, chemical, and physical attributes.
- Developed a Logistic Regression model to evaluate anti-HBV properties of compounds, leveraging a dataset of 50k pre-analyzed molecules, and achieving a ROC_AUC score of 0.82.
- Identified 6 potential drug candidates from 1 M screening pool, obtaining a six-fold increase in discovery efficiency.

Genetic Mutant Identification for Alcohol Addiction:

- Utilized generalized linear mixed-effect models to analysis RNA sequencing results from alcohol addiction patients, enabling precise detection of mutations associated to alcohol exposure.
- Implemented Random Forest to classify 1 M in silico mutant candidates, achieving a precision of 0.5 and a recall of 0.95 through iterative model refinement.
- Identified 280 key causal mutants for further analysis, which significantly streamlined research direction and cut down experimental costs by 60%.

EDUCATION

Chinese Academy of Sciences	Ph.D. in Microbiology	Beijing, China	2008
Purdue University	Master’s in Applied Statistics	Indiana, US	2022