CHUNLIN HE, PhD

Data Scientist/Machine Learning Scientist

chhe09@gmail.com • (765) 490-7331 • West Lafayette, IN • https://datute.net

SUMMARY

- Results-driven Data Scientist with a Ph.D. and MSc in Biostatistics, bringing 14+ years of experience in quantitative and qualitative research, predictive analytics, statistical modeling, machine learning, and deep learning. Skilled in leveraging data to uncover business insights, identify opportunities, and support data-driven decision-making.
- Proficient in Python, R, SQL, and advanced AI techniques, including NLP, LLMs, and RAG, with expertise in applying statistical, algorithmic, and data mining methods as well as advanced visualization techniques. Experienced in model risk management, performance optimization, and end-to-end project delivery.
- Successfully led multi-million-dollar data initiatives, including genomic data analysis across 50+ countries, and provided actionable insights for clients in the finance and insurance sectors. Recognized for driving digital transformation, building scalable AI/ML solutions with UI integration, and effectively presenting complex findings to technical and non-technical audiences.
- Published author of 58 peer-reviewed papers and a trusted collaborator known for leadership, innovation, and delivering measurable impact.

TECHNICAL SKILLS

Category	**Skills/Techniques**
Programming Languages	Python, R, SQL, SAS, JavaScript.
Libraries	NumPy, Pandas, Matplotlib, Seaborn, TensorFlow, scikit-learn, SciPy, Keras, PyTorch, ggplot2, Caret, tidyverse.
Data wrangling and statistical analysis	EDA, Hypothesis Testing (parametric and non-parametric), Correlation and Regression Analysis, Chi-square Test, ANOVA, t-test (including independent and paired samples), A/B Testing, Statistical Modeling (linear and nonlinear), PCA, Factor Analysis, Time Series Analysis, Cluster Analysis, Bayesian Statistics.
Cloud Platforms	AWS (Sagemaker, Lambda, S3, Redshift, Glue, Rekognition, EMR, Data Pipeline, CloudWatch), Google Cloud (VertexAI, BigQuery), MS Azure (Databricks).
Machine Learning and Deep Learning	Linear Regression, Logistic Regression, Random Forest, Decision Tree, SVM, Gradient Boosting (AdaBoost, XGBoost, CatBoost), KNN, Neural Networks, NLP, LLMs and RAG, computer vision.
MLOps	AWS Sagemaker, Docker, Jenkins, Kubernetes, Ansible, AWS Batch, VertexAI, Azure ML, Data Factory, DevOps, AKS

Other Skills	GitHub, PySpark, GCP BigQuery, MS SQL Server, MySQL, SQLite, RStudio, Jupyter, Linux, Anaconda, Big Data, Hadoop, NLTK, Power BI, Tableau, REDcap, AWS, Web development.

EXPERIENCE

Outlier AI, West Lafayette, IN

Data Scientist/Machine Learning Scientist/AI Trainer

Aug 2023 - Present

- Extracted, cleaned, and preprocessed large datasets from various sources to uncover insights and facilitate analysis using Python libraries such as Pandas, NumPy, Scikit-learn, and PySpark.
- Conducted statistical analysis and hypothesis testing on complex business data to derive data driven insights and inform strategic decision-making processes.
- Developed and implemented machine learning models for predictions, leveraging different datasets to fine-tune different classification and regression models.
- Utilized Google Cloud Platform (GCP) services, including VertexAI, BigQuery and Cloud SQL, to manipulate and analyze data efficiently at scale.
- Translated data insights into actionable recommendations, driving business outcomes and strategic initiatives.
- Communicated findings and recommendations effectively to stakeholders through clear and concise data visualizations and presentations.
- Fine-tuned and evaluated large language models (#LLMs & #RAG), identified and mitigated inherent biases to optimize performance across diverse tasks and domains.

TCS/USAA, West Lafayette, IN

Data Scientist and Modeler

Sep 2020 - Jul 2023

- Developed and deployed predictive models for classification and regression using Python (NumPy, Pandas, scikit-learn, PySpark) and applied MLOps practices, driving measurable improvements in business operations across the insurance and finance sectors.
- Applied quantitative and qualitative research methods, including sentiment analysis and hypothesis testing, to enhance model risk management, uncover business insights, and improve customer satisfaction and retention.

- Conducted rigorous evaluations of machine learning models, implementing statistical, algorithmic, and data mining methods to optimize parameters, improve predictive accuracy, and ensure alignment with client objectives in property and casualty (P&C) insurance.
- Leveraged advanced visualization techniques to communicate model performance, weaknesses, and opportunities, enabling stakeholders to make informed, data-driven decisions.
- Provided expert guidance throughout the model development lifecycle, from survey data utilization and validation to deployment and maintenance, ensuring sustained performance and continuous business enhancement.

Microsoft Professional Program, West Lafayette, IN

Freelance Data Scientist

Jan 2019 - Aug 2020

- Led data mining, wrangling, cleansing, integration, and visualization for large-scale health and clinic datasets (12M+ records), ensuring data integrity and actionable insights using SQL, Power BI, Tableau, and REDCap.
- Applied quantitative statistical methods (R, dplyr, tidyr, ggplot2) to extract and visualize trends from health research data, enabling evidence-based decision-making.
- Designed and implemented a relational database management system (RDBMS) in MS SQL Server for a retail company, improving data manipulation, storage, and retrieval efficiency.
- Executed machine learning projects (classification, regression, and forecasting) with Python (scikit-learn, PySpark, Prophet), including predicting mortgage rate spreads, forest cover types, customer spending, and COVID-19 time series forecasting.
- Applied computer vision techniques (TensorFlow, PyTorch) in the Global Wheat Head Detection project, accurately detecting wheat heads for yield estimation across diverse environments.
- Consistently combined statistical, algorithmic, and qualitative insights to deliver predictive models, uncover business opportunities, and enhance research and operational outcomes.

Noble Research Institute, Ardmore, OK

Alfalfa Data Curator

Jun 2015 - Dec 2018

- Led data analysis and curation for the Alfalfa Toolbox project, accelerating molecular breeding for the 8-billion-dollar alfalfa market.

- Applied R packages, dplyr and tidyr, for data manipulation, transformation and statistical analyses for numerous datasets of field experiments including using ggplot2 package for versatile data visualization. Conducted hypothesis testing to validate experimental results.
- Applied Python libraries, including NumPy, Pandas, scikit-learn (SKLearn), SciPy, and StatsModels, for statistical analysis of multiple datasets from alfalfa populations. Created statistical models and used Matplotlib and Seaborn for data visualization.
- Developed predictive models using machine learning algorithms, including Linear Regression, Logistic Regression, Random Forest and SVM to predict the potentials of important traits from elite alfalfa populations that were selected for curation in the toolbox.
- Maintained ETL pipelines and databases using RDBMS (MS SQL Server) to efficiently extract, transform, and load alfalfa population data with desired traits for integration into the Alfalfa Toolbox project.
- Optimized data integration into the Alfalfa Toolbox web portal by streamlining processes through the Toolbox API, ensuring seamless access and usability.
- Developed UI components and collaborated on interactive data visualization dashboards for the Alfalfa Toolbox web portal using HTML, CSS, JavaScript, D3.js, and Plotly, enhancing user experience and accessibility

International Maize & Wheat Improvement Center, Mexico City, MX

Senior Scientist and Project Manager (Genomic Data)

Oct 2010 – Dec 2014

- Managed over 100 genomic data projects, overseeing the generation of genomic and genotypic data for academic and research institutions across 50+ developing countries.
- Applied Bioconductor R packages to preprocess raw data obtained from high-throughput genomic experiments including quality control and normalization for downstream analysis, such as clustering and classification of datasets.
- Performed Genome-wide Association Studies (GWAS) using Bioconductor R packages, involving analyzing genetic variants across the genome to identify associations with traits or diseases, including association testing and visualization of GWAS results.
- Used the R/qtl package to identify QTLs and estimate genetic distances for multiple economic traits in the elite crosses of different legume crops.
- Collaborated with research scientists to analyze genetic and genomic data using various tools including MapMaker, JoinMap, Mapchart, MapDisto, SAS PROC QTL, R-GWLD, R and Bioconductor for genomic analysis, including genetic mapping, QTL analysis & data visualization.

- Mentored and guided a team of research scientists, leveraging expertise in cutting-edge data analytics tools and visualization techniques for the effective analysis of genotypic and phenotypic data using advanced statistical modeling techniques and Causal Inference Modeling.

EDUCATION

University of Saskatchewan — *Doctor of Philosophy - PhD, Plant Science focusing on Biostatistics*

Nanjing Agricultural University — *Master's degree in Biostatistics and Quantitative Genetics*

Hunan Agricultural University — *Bachelor's in Crop Science*

CERTIFICATIONS

Databricks — *Professional Certificate, Large Language Models*

Microsoft — *Professional Certificate in Data Science*

Microsoft — *Data Science Certification, Microsoft Professional Program (MPP)*