CHUNLIN HE, PhD

Data Scientist/Machine Learning Scientist

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

SUMMARY

Experienced Data Scientist with a Ph.D. and MSc in Biostatistics, backed by 14+ years of expertise in data generation, statistical analysis, machine learning, and deep learning techniques. Proficient in Python, R, SQL, and more, accomplished at leading teams and independently driving projects from conception to deployment.

Skilled in curating and analyzing complex datasets, demonstrated through successful management of multi-million-dollar data projects in the research sectors, including contributions to the plant genomic data analysis initiatives across 50+ countries.

Strong background in digital transformation and rapid prototyping, with a focus on developing actionable machine learning solutions and UI components using advanced analytics methods including NLP, LLMs and RAG (Model Training/Research). Proven ability to identify performance improvement opportunities, conduct Proof of Concepts, and communicate complex technical findings effectively to diverse stakeholders.

Experienced in collaborating with clients in the finance and insurance sectors, contributing to model risk management and enhancing business performance. Recognized for leadership, communication, and organization skills, with a track record of delivering results and driving innovation.

Contributed technical expertise, leadership, and problem-solving skills to drive meaningful impact within a dynamic and innovative organization.

Accomplished author and co-author of 58 peer-reviewed papers, conference papers, and book chapters related to biostatistics and data analytics.

TECHNICAL SKILLS

Category	Skills/Techniques
Programming Languages	Python, R, SQL, SAS, JavaScript
Libraries	NumPy, Pandas, Tensorflow, scikit-learn, SciPy, Keras, PyTorch, ggplot2, Caret, tidyverse, KQL
Data wrangling and statistical analysis techniques:	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, EKS), Google Cloud (VertexAI, BigQuery), 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.
MLOps	AWS Sagemaker, Docker, Jenkins, Kubernetes, Ansible, AWS Batch, VertexAI.
Other Skills	Github, Azure ML, PySpark, GCP BigQuery, MS SQL Server, MySQL, SQLite, RStudio, Jupyter, Linux, Anaconda, Big Data, Hadoop, NLTK, Matplotlib, Power BI, Tableau, REDcap, AWS, Web development.

EXPERIENCE

Outlier AI/Databricks/DataAnnotation, West Lafayette, IN — *Data Scientist/Machine Learning Scientist/LLM Trainer & Researcher*

August 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/Modeler*

September 2020 - July 2023

- Specialized in developing models using Python libraries such as NumPy, Pandas, scikit-learn, PySpark, and implementing MLOps practices to tackle classification and regression problems in business operations within the insurance and finance sectors.
- Utilized sentiment analysis to develop and assess predictive models, enhancing model risk management and business performance with a focus on improving customer satisfaction and retention. Performed hypothesis testing to ensure the reliability of sentiment analysis results.
- Engaged in comprehensive evaluation of machine learning models, conducting rigorous quality assurance and quality control processes to optimize model parameters and enhance predictive accuracy, with a focus on clients in insurance and finance.
- Evaluated model accuracy and weaknesses through rigorous testing, driving performance improvements and ensuring alignment with business objectives.

- Provided extensive guidance throughout the model development process, offering insights on utilizing custom survey data effectively and ensuring robust validation techniques to enhance predictive capabilities and customer satisfaction.
- Implemented effective model maintenance practices, contributing to continuous business enhancement and ensuring sustained performance.

Microsoft/MPP, West Lafayette, IN — *Freelance Data Scientist*

January 2019 - August 2020

- Managed data mining and database operations for a private health company (NDA), ensuring data integrity and accuracy. Applied REDcap for data collection, management, cleansing, integration and visualization using Power BI and Tableau.
- Utilized complex SQL queries to work with large datasets (over 12 million observations of health and clinic data), focusing on quality assurance, data extraction, and merging tasks from multiple sources.
- Analyzed health datasets from a private research organization using R packages, dplyr and tidyr, providing valuable insights derived from statistical analysis and data visualization using ggplot2.
- Developed and implemented a relational database management system (RDBMS) using MS SQL Server to create an inventory database for a retail company. Successfully executed data manipulation, storage, and retrieval operations.
- Conducted machine learning project as part of Microsoft's MPP program, specializing in classification and regression including mortgage rate spreads across 50 states.
- Executed machine learning projects (regression and classification) using Python, Scikit-Learn, Pandas, and PySpark to predict forest cover types and customer budget spending through various algorithms, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBClassifier.
- Developed a time series model using Python library, Prophet, for forecasting the covid-19 pandemic in 2020 that was caused by the pneumonia-causing novel coronavirus (SARS-CoV-2).
- Completed the Global Wheat Head Detection project using deep learning models (using TensorFlow or PyTorch frameworks) to accurately detect wheat heads in outdoor field images, enabling the estimation of wheat head density and size across diverse growing environments worldwide

Noble Research Institute, Ardmore, OK — *Data Curator/Data Analyst*

June 2015 - December 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 and Wheat Improvement Center, Mexico City, MX — *Senior Scientist and Project Manager (Genomic Data)*

October 2010 - December 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 with focus 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)*