CHUNLIN HE, PhD

Data Scientist/Machine Learning Scientist

chhe09@gmail.com • (765) 490-7331 • West Lafayette, IN • LinkedIn • Github • Website

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 using advanced analytics methods including NLP and LLMs. 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.
- Contributing technical expertise, leadership, and problem-solving skills to drive meaningful impact within a dynamic and innovative organization.

CERTIFICATION & EDUCATION

- Professional Certificate – Large Language Models, Databricks, USA.
- Data Science Certificate: Microsoft Professional Program (MPP) for Data Science, Microsoft, USA
- Professional Certificate: Data Science Fundamentals, Microsoft, USA.
- Ph.D.: Plant Science with a focus on Biostatistics, University of Saskatchewan, Canada.
- M.Sc.: Biostatistics and Quantitative Genetics, Nanjing Agricultural University, Nanjing, China.
- B.Sc.: Crop Science, Hunan Agricultural University, Changsha, China.

SKILLS

- **Programming languages:** Python (with TensorFlow, Keras, Scikit-learn), R (with ggplot2, caret, tidyverse), SQL (including KQL), SAS, and JavaScript.
- **Machine Learning (ML) and Deep Learning (DL):** Linear Regression, Logistic Regression, Random Forest, Decision Tree, SVM, Gradient Boosting (AdaBoost, XGBoost, CatBoost), KNN, Neural Networks, NLP, and LLMs.

- **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
- **Platforms, databases and IT tools:** Github, Azure ML, PySpark, GCP BigQuery, MS SQL Server, MySQL, SQLite, RStudio, Jupyter, Linux, Anaconda, Big Data, Hadoop, NLTK, Matplotlib, AWS, Web development.

PROFESSIONAL EXPERIENCE

Freelance Data Scientist/Machine Learning Scientist, West Lafayette, IN **Aug 2023 – present**

Roles and Responsibilities:

- Extracted, cleaned, and preprocessed large datasets from various sources to uncover insights and facilitate analysis.
- 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 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 LLMs, identifying and mitigating inherent biases to optimize performance across diverse tasks and domains.

Modeller/Data Scientist, Tata Consultancy Services Ltd., West Lafayette, IN **Sep 2020 – Jul 2023**

Roles and Responsibilities:

- Utilized quantitative analysis to assess predictive models, enhancing model risk management and business performance, particularly within the insurance and finance sectors.
- 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.
- Specialized in developing models for business operations, with an emphasis on improving customer satisfaction and retention.
- 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 Data Science Track & Freelance Data Scientist, W. Lafayette, IN Jan 2019 – Aug 2020

Roles and Responsibilities:

- Managed data mining and database operations for a private health company (NDA), ensuring data integrity and accuracy. Implemented data cleansing and query operations for large datasets.
- Utilized SQL queries (using MS SQL Server and SQLite) to work with large datasets (over 12 million observations of health and clinic data), focusing on quality assurance/control, data extraction, and merging tasks.
- 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 the machine learning projects to predict the forest cover types and customer budget spending using different algorithms including (1) Logistic Regression, (2) Support Vector Machine (SVM), (3) Random Forest, (4) 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.

Alfalfa Toolbox/Data Curator, Noble Research Institute, Ardmore, OK May 2015 – Dec 2018

Roles and Responsibilities:

- 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.
- Applied Python libraries, including NumPy, Pandas, SciPy and StatsModels, for statistical analysis for multiple datasets from alfalfa populations as well as creating statistical models, and using 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 tool box.

- Maintained databases using RDBMS (MS SQL Server), specializing in efficient extraction of alfalfa populations with desired traits for users, partners and customers in 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.

Breeding Services Manager (Genomic Data)/Senior Scientist, GCP/CIMMYT, Texcoco, Mexico

Oct 2010 – Dec 2014

Roles and Responsibilities:

- 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.

REFERENCES

Available upon request.