CHUNLIN HE, PhD

Data Scientist/ Machine Learning Scientist

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

SUMMARY

Results-oriented Data Scientist with a Ph.D. and an MSc in Biostatistics, and professional certificates in Data Science and LLMs. Proficient in Python, R, SQL, SAS, with expertise in Machine Learning and Deep Learning, including NLP and LLMs. Skilled in developing and optimizing predictive models, conducting thorough assessments, and refining models for improved business performance. Proven experience in managing end-to-end data projects, collaborating on health industry initiatives, clinic data analysis, and implementing relational database management systems. Experienced in working with clients in the insurance and finance sectors, contributing to model risk management and enhancing business performance. Basic understanding of AWS services through training. Accomplished in multiple statistical and quantitative research projects, with a notable portfolio of 58 peer-reviewed papers and book chapters. A dynamic problem-solver poised to excel in a Data Scientist role.

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, R, SQL, KQL, SAS, 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 & LLMs.
- Data wrangling and statistical analysis: EDA, Hypothesis Testing, Correlation and Regression Analysis, Chi-square Test, ANOVA, t-test, A/B Testing, Statistical Modeling, Principal Component Analysis, Nonparametric Statistics.
- Tools and Technologies: Github, AzureML, Power BI, Spark, Tableau, HDInsight, AWS, MS SQL Server, MySQL, SQLite, RStudio, Kusto.Explorer, Google Colab, GCP BigQuery, AI Platform.

PROFESSIONAL EXPERIENCE

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

- Extracted, cleaned, and preprocessed large datasets to uncover insights and facilitate analysis.
- Analyzed complex business data to discover trends and patterns for strategic decision-making.
- Conducted statistical analysis and hypothesis testing to provide data-driven insights.
- Develop and implement machine learning models for predictions, including the use of GCP platform with Big Query and Cloud SQL
- Fine-tuned and evaluated LLMs including identifying and mitigating inherent biases to optimize performance across diverse tasks.

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

- Utilized quantitative analysis to assess predictive models, enhancing model risk management and business performance, particularly with clients in the insurance and finance sectors.
- Specialized in developing models for business operations, focusing on customer satisfaction and retention.
- Evaluated model accuracy and weaknesses through rigorous testing, driving predictive performance improvements.
- Implemented effective model maintenance practices, contributing to continuous business enhancement.

Microsoft/MPP Data Science Track & Freelance Data Scientist, W. Lafayette, IN Jan 2019 – Aug 2020

- Managed data mining and database operations for a private health company (NDA), ensuring data integrity and accuracy. Implemented data cleansing, query operations, and merging processes for large datasets.
- Analyzed health datasets from a private research organization using R, providing valuable insights derived from statistical analysis.
- Developed and implemented a relational database management system (RDBMS) using MS SQL Server. Executed data manipulation, storage, and retrieval operations effectively.
- Completed machine learning projects as part of Microsoft's MPP program, specializing in classification and regression. Addressed real-world data science challenges encountered by research organizations.

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

- Led data analysis and curation for the Alfalfa Toolbox project, accelerating molecular breeding for the 8-billion-dollar alfalfa market.
- Managed end-to-end data pipeline and conducted exploratory data analysis (EDA) using Python and R on diverse datasets from multi-environmental trials, deriving actionable insights.
- 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

- Managed over 100 genomic data projects, overseeing the generation of genomic and genotypic data for academic and research institutions across 50+ developing countries.
- Applied advanced data analysis techniques such as linkage analysis and QTL mapping to extract actionable insights, facilitating the integration of genotypes and phenotypes for marker-assisted breeding.
- 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.