CHUNLIN HE, PhD
Data Scientist, Microsoft Certified

Cell: 765-490-7331 • chhe09@gmail.com • https://datute.net

SUMMARY

- Data scientist proficient in Python, utilizing packages such as Pandas, NumPy, Matplotlib, SciPy, Seaborn, SciKit-Learn, Statsmodels, TensorFlow, and more.
- Experienced in developing classification and regression models using machine learning algorithms like Logistic Regression, k-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Regression (SVR), Linear Regression, and Gradient Boosting.
- Evaluated predictive models for both banks and non-bank institutions. Identified gaps for model enhancement and risk management, enabling data-driven insights and improved customer services and membership retention.
- Managed databases and led data analysis for the alfalfa toolbox project (ABT), a multi-million dollar initiative aimed at accelerating molecular breeding in the 8-billion dollar alfalfa market.
- Successfully managed over 100 data projects, collaborating with 100+ academic and research institutions across 50+ developing countries to generate and analyze genotypic data.
- Demonstrated proficiency in additional data science tools, including SQL, Bash, SAS, MS SQL Server, MySQL, NoSQL (MongoDB), Jupyter Notebook, Google Colab, RStudio, AzureML, Power BI, Tableau, and Big Data analytics with MS R Client, HDInsight, and Spark.
- Authored/co-authored 58 peer-reviewed & conference papers including book chapters.

EDUCATION

- 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 AND KNOWLEDGE

- Computing skills: Python, R, MS R Client, SAS, Minitab, SQL, Git Bash, JavaScript, AzureML, Power BI, Spark, Tableau, HDInsight, MS SQL Server, MySQL, MongoDB, HTML, CSS, Google Colab, Jupyter Notebook.
- Machine Learning (ML) and Deep Learning (DL): Linear Regression, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, Gradient Boosting, AdaBoost, XGBoost, CatBoost, Neural networks, computer vision, image analysis, model assessment.
- Data wrangling and statistical analysis: Data pipeline, data wrangling, exploratory data analysis (EDA), hypothesis testing, correlation and regression analysis, Chi-square test, ANOVA, t-test, statistical modeling, principal component analysis, nonparametric statistics.

- Bioinformatics: Genome sequence assembly, annotation, DNA sequence analysis, gene identification, RNA-seq, gene expression analysis, sequence motif, SNP markers, DNASTAR, GeneMapper, MapQTL, QTL IciMapping, Biopython.

PROFESSIONAL EXPERIENCE

Modeller/Data Scientist, Tata Consultancy Services Ltd. | 2020 - present

- Assessed predictive bank and non-bank models as a Data Scientist, identifying gaps for model risk management to provide insights and improve business performance for clients.
- Focused on models related to business operations and performance, specifically predicting customer satisfaction and membership retention.
- Played a vital role in conducting exploratory data analysis, data wrangling, variable selection, model development assumptions, data analysis, testing, benchmarking, and evaluating the accuracy, weaknesses, and limitations of models.
- Implemented model maintenance, monitoring, and documentation to drive business performance improvements.

Microsoft/MPP Data Science Track & Freelance Data Scientist | 2019 - 2020

- Data mining and database management: Collaborated with a private company in the health industry under a non-disclosure agreement (NDA) to perform data cleansing, query and merging operations, and explore relationships within the dataset. Managed and maintained the database effectively to ensure data integrity and accuracy.
- Clinic data analysis: Conducted statistical analysis using R on a comprehensive dataset obtained from a private research organization (NDA). Analyzed data sets aggregated from tens of thousands of participants, providing valuable insights to support decision-making and enhance organizational outcomes.
- Relational database management project: Successfully completed a project for a private company involving the implementation of a relational database management system (RDBMS) using MS SQL Server. Designed and developed a robust database solution, optimizing data storage, retrieval, and management processes.
- Microsoft/MPP machine learning projects: Demonstrated expertise by successfully completing various machine learning projects, addressing real-world data science challenges commonly encountered by corporations and research organizations.

Alfalfa Toolbox/Data Curator, Noble Research Institute, Ardmore, OK | 2015 - 2018

- Curated and analyzed alfalfa data for the multi-million dollar Alfalfa Toolbox project (ABT), aimed at accelerating molecular breeding for the 8-billion dollar alfalfa market.
- Conducted data pipeline and exploratory data analysis (EDA) using Python and R on diverse datasets from multi-environmental trials, generating valuable insights.
- Designed and maintained databases using RDBMS (MS SQL Server), efficiently extracting populations of alfalfa with multiple traits of interest for the ABT, partners, and customers.
- Optimized data integration into the ABT web portal through the Toolbox API.

Breeding Services Manager/Senior Scientist, GCP, c/o CIMMYT, Mexico | 2010 - 2014

- Oversaw and managed 100+ data projects, generating genomic and genotypic data for more than 100 academic and research institutions across 50+ developing countries.
- Utilized data analysis techniques, including QTL mapping, to derive valuable insights bridging genotypes and phenotypes for international breeding programs.
- Mentored and guided research scientists, demonstrating data analytics tools and visualizations for effective analysis of genotypic and phenotypic data.