# Hao Zhou

Address: New Orleans, LA 70112 || Cell: (504) 300-3134 || 952774178z@gmail.com
LinkedIn: shorturl.at/cyBL0

## SUMMARY

Master student seeking data-related opportunities, with strong background in **mathematics**, **statistics**, and sufficient experience in **machine learning** and **data analytics**. Solid programming skills in **Python**, **SQL**, and a keen interest in deep learning.

## EDUCATION

**Tulane University** *New Orleans, LA* Expected May 2025
Master of Science in Biostatistics
**Nanjing University** *Nanjing, China* June 2022
Bachelor of Science in Statistics

## SKILLS

**Programming and Tools:** Python, SQL, SPSS, Spark, PyTorch, Tableau

**Machine Learning**

- Classical Linear Models, Decision Tree, Random Forest, Gradient Boosting Decision Tree, K Nearest Neighbors (KNN), Support Vector Machines (SVM), Recommendation System, Neural Network
- Clustering, Natural Language Processing (NLP), Latent Dirichlet Allocation (LDA)
- Exploratory Data Analysis, Data Preprocessing, Data Visualization, Principal Component Analysis (PCA), Regularization, Feature Engineering, Model Evaluation

**Statistics Analysis:** Parameter Estimation, Hypothesis Testing, Analysis of Variance (ANOVA), Time Series

## PROJECTS

**Bird Image Classification using Convolutional Neural Network**

- Built a Convolutional Neural Network model using PyTorch to classify 90K images spanning 525 bird species.
- Implemented Data Augmentation and Data Normalization.
- Utilized pretrained models including AlexNet and ResNet, fine-tuning them to adapt to the output requirements for the classification task.
- Evaluated the model on test data and identified the best model with an accuracy of 98%.

**Anime Recommendation System based on Matrix Factorization**

- Implemented a recommendation model in PySpark to suggest anime likely to interest users.
- Conducted Online Analytical Processing (OLAP) with Spark SQL.
- Trained Alternating Least Squares (ALS) model for Matrix Factorization to provide personalized recommendations, and developed user-based and item-based algorithms to handle cold start problems.
- Tuned model hyperparameters through cross validation and achieved the optimal model with an RMSE of 1.2 in a user rating range of 1-10.

**Natural Language Processing on ChatGPT Prompts**

- Clustered ChatGPT prompts into groups, unveiling common scenarios where users engage with the model.

- Preprocessed the text data through tokenization, stemming, and TFIDF feature extraction.
- Employed K-means clustering and Latent Dirichlet Analysis (LDA) for unsupervised learning, identifying different prompt patterns and topics.
- Applied t-SNE for dimensionality reduction and visualized clustering results.

**Employment Prediction for Graduates from a Campus**

- Developed a model in Python to predict whether students would get placed after graduation based on grades, degree specialization and work experience.
- Conducted comprehensive data preprocessing, including data cleaning, categorical feature transformation, and data standardization.
- Established Logistic Regression, Random Forest, KNN and SVM models, and used Grid Search based on 5-fold cross validation to find optimal model hyperparameters.
- Evaluated the models on test data and selected the best model based on AUC (best AUC: 0.96).

**Stock Price Forecast Based on Time Series Analysis**

- Developed a model in Python to predict stock prices for the upcoming week.
- Processed over four years of historical data and established dummy variables to account for the impact of the COVID-19 pandemic.
- Trained models including ARIMA and XGBoost to forecast stock price for a week.
- Evaluated model performance on test data with a minimal RMSE of 2.7.

**California Housing Price Prediction**

- Developed algorithms in Python for predicting California housing prices and determined the importance of each feature in the model.
- Created interactive maps visualizing housing quantities and prices across different regions in California.
- Implemented feature engineering based on geospatial analysis and trained models such as Random Forest, Gradient Boosting, and Multilayer Perceptron (MLP) to predict housing prices.
- Evaluated model performance and analyzed feature importance to identify the top 5 factors influencing housing prices.