Xiao Song

songxiao@umd.edu songxxiao xsong.ltd.en | *Updated:* July 2, 2020

Education

Bachelor of Sociology East China Normal University 2016~2020

Academic Research

Machine Learning in Social Sciences: Based on China Education Panel Survey 2019~2020

Bachelor Degree Thesis (PDF)

Welfare Effect and Social Inequality of Land Transfer: Empirical Analysis Based on CFPS 2018~2019

The data of China Family Panel Survey (CFPS) were used for data cleaning and econometric analysis through Stata and R. Using Unconditional Quantile Regression and Fixed Effect Model estimate the welfare effect of land transfer behavior and its impact on social inequality. Using R's `ggplot2` software package to visualize geographic information

Work & Internship Experience

Remote Data Scientist Internship 2020-02~2020-04
Zhongnan University of Economics and Law Data Consultant
Use Xgboost, RandomForest, LightGBM and other algorithms to classify (multiclass) legal text data. The word frequency method is used to construct the feature matrix, and the cross-validation training model (`sklearn`) is used to obtain the cross-validation accuracy of 0.75. I write a program to make predictions on new data, so that the prediction results can be applied to any new data set.

Data Analyst 2019-07~2019-09
iResearch Using R and SPSS to analysis profile of cars' users. Through PCA and Cluster analysis, I catogorized survey data and found cars users' attitude difference. Using MySQL database to help analyze users' data. Using Hive SQL to help access Hadoop database.

Awards and Honors

Kaggle M5 Forecasting - Accuracy 103rd/5558 Top2% Silver Medal 2020
Estimate the unit sales of Walmart retail goods

Kaggle M5 Forecasting - Uncertainty 18th/909 Top2% Silver Medal 2020
Estimate the uncertainty distribution of Walmart unit sales

Skills

Data Analytics
Familiar with the principle and implementation of statistical analysis in R language, able to use `tidyverse`, `data.table` for data cleaning.
Understand R language statistical analysis, derivation and implementation of LR, RNN, generalized linear model, K-means and other methods.
Familiar with Python `pandas` library for manipulation of tabular data, `numpy` library for numerical operations.

Data visualization
Familiar with R language `ggplot2`, Python `seaborn` `plotnine` library.

Machine Learning
Understand the principles and implementation of Xgboost, LightGBM and other high-performance algorithms.
Familiar with the implementation of various algorithms of Python `sklearn` library, including supervised learning and unsupervised learning.
Understand the principles and implementation of natural language processing using `keras` deep learning framework.

Big Data
Understand how to use R language `sparklyr` and other tools to connect and operate Hadoop and Spark clusters

Other Skills
SQL, SPSS, Stata, Git, LaTeX, MS Office, HTML/CSS

Standard Examination
TOEFL 103 *Reading* 29 *Listening* 27 *Speaking* 21 *Writing* 26
GRE *Verbal* 154 *Quantity* 167 *Writing* 3.5