宋 骁

期望职位：数据分析、数据科学

邮箱：songxiao@umd.edu ◇ 个人网站：https://xsong.ltd/

教育背景

华东师范大学,社会发展学院, 社会学 *2016 年-2020 年*
相关课程：R 语言数据分析、算法与程序设计基础 (Python)、回归分析与 Stata 应用
SPSS 应用、高级统计与论文写作、社会模拟与 NetLogo 应用

北京大学-密歇根大学学院, 北京大学, 暑期课程 *2018 年 8 月*
因果推论方法的研究设计和敏感性分析

工作经历

中南财经政法大学　法律文本分类 *2020 年 2 月-2020 年 3 月*

- 使用 Python 的 `pandas`、`numpy` 和 `scikit-learn` 库进行数据清理、建模。使用决策树和随机森林算法建立信用卡评分模型。通过交叉验证 (Cross-Validation) 的方法训练模型，对互联网法律文本数据进行分类。

艾瑞咨询　数据分析实习生 *2019 年 7 月-2019 年 9 月*

- 使用 Hive SQL 对 Hadoop 业务数据库集群进行查询和清理
- 使用 R、Python 对文本数据进行可视化

学术研究

机器学习在社会科学实证研究中的应用 *2020 年　学士学位论文*

- 独立作者。指出基于广义线性的传统定量分析方法具有模型的不确定性问题，研究者也存在滥用 P 值的不规范行为，这导致研究结果的不可靠性与不可重复的问题。机器学习中的集成学习方法组合多个学习器，在保持模型稳定性的同时具有更强大的预测能力与可解释性，它们能够对现有统计方法进行补充并得到广泛应用。

土地流转的福利效应与社会不平等 *2019 年　国家大学生创新训练项目*

- 独立作者，使用中国家庭追踪调查 (CFPS)数据，通过 Stata 和 R 进行数据清理和计量经济分析。使用无条件分位数回归和固定效应模型估计土地流转行为的福利效应和对社会不平等的影响。

获奖与资助

银牌 (103rd/5558, Top2%)　M5 数据挖掘竞赛 沃尔玛销量时序预测-准确性　*2020 年 Kaggle 主办*

银牌 (18th/909, Top2%)　M5 数据挖掘竞赛 沃尔玛销量时序预测-不确定性分布　*2020 年 Kaggle 主办*

优秀奖　第二届全国高校数据驱动创新研究大赛　*2019 年 北京大学图书馆主办*

二等奖　优秀学生奖学金　*2017 年，2018 年 华东师范大学*

核心技能

数据处理	熟悉 MySQL 数据库语言。能够使用 R 语言与 SQL 进行连接和操纵，提升分析性能
	熟悉 R 语言统计分析的原理与实现，能够使用 tidyverse,data.table 进行数据清洗
	熟悉 Python pandas 库对表格数据的操纵，numpy 库进行数值运算
	了解 R 语言统计分析，对 LR，RNN，广义线性模型，K-means 等方法的推导及实现
数据可视化	熟悉 R 语言 ggplot2，Python seaborn plotnine 库
机器学习	熟悉 Python sklearn 库的各算法实现，包括监督学习和无监督学习
	了解 Xgboost, LightGBM 等高性能算法的原理与实现
	了解使用 Keras 深度学习框架自然语言处理的原理与实现
大数据分析	了解使用 R 语言 sparklyr 等工具连接并操作 Hadoop，Spark 集群
其他技能	SPSS, Stata, Git, LaTeX, MS Office, HTML/CSS
标化考试	大学英语四级 614，六级 549，托福 103(阅读 29　听力 27　口语 21　写作 26),
	GRE 321(语文 154　数学 167　作文 3.5)