唐雨辰

yt2754@columbia.edu 19801285571 LinkedIn

教育经历

哥伦比亚大学 | 纽约，美国 **2021.01-2022.05**
统计学，硕士学位 GPA: 3.97/4.0
核心课程: 统计推断，线性回归模型，时间序列模型，抽样调查，数据科学

国际关系学院 | 北京，中国 **2015.09-2019.06**
信息管理与信息系统，学士学位 GPA: 3.73/4.0
核心课程: C 语言，数据结构与算法，数据库原理，数据仓库与数据挖掘

北京大学，国家发展研究院 | 北京，中国 **2016.09-2019.06**
经济学，双学士学位 GPA: 3.47/4.0
核心课程: 中级宏/微观经济学，计量经济学，博弈论

技能

硬技能 MS Office 套装, R 语言 == 统计学 > SQL > Python, Jupyter Notebook, (R)Markdown, Latex, Tidyverse, RShiny, 数据可视化 (ggplot2, seaborn), 假设检验, 机器学习模型: 线性回归, Logistic 回归, 基于树的模型, 神经网络, 聚类

语言技能 英语四级 660, 英语六级 570, 英语专业四级 合格, 英语专业八级 合格, 托福 104, GRE 322

软技能 自我学习，多任务处理，以解决方案导向

实习经历

助教 | 统计系，哥伦比亚大学 | 纽约，美国 **2021.09-2022.05**
- 修订和批改作业，为本科生和研究生进行统计学和 R 语言编程习题的答疑

数据源管理实习生 | 企查查科技有限公司 | 苏州，中国 **2020.07-2020.10**
- 参与设计了一个招投标信息的新板块，包括确定新板块的结构布局和所需要的功能
- 手动收集了 300+的数据源；使用 SQL 查询来查找不确定和缺失的数据，与爬虫小组进行对接来反馈错误

助理分析师 | TokenInsight 未来通证科技有限公司 | 北京，中国 **2019.04-2019.07**
- 为了估计加密货币交易所的虚假交易量 (wash trading) 及真实交易量，提出幂律分布 (Power Law) 可以被用来检验交易所订单的分布，然后识别其中的反常表现，并进行矫正
- 分别使用 R 语言和 Python 实现了这个设想. 使用原始的订单簿数据验证了它的有效性，其过程包括数据处理 (分段，极端值处理)，可视化 (直方图，散点图)，曲线拟合 (对数变换，线性回归) 等
- 分析了 10+交易所，发现了它们的洗牌交易的特征并估计了其虚假和真实的交易量；完成了最终报告中的相应部分；公司成为国内最早发布使用定量方法估计交易所真实交易量报告的机构

项目经历

有关幸福感的研究 **2022**
- 使用 Python 预处理数据，例如 one-hot/label 编码，数据标准化，特征整合等特征工程；使用 scikit-learn 包的 logistic 回归和梯度提升树进行建模，使用网格搜索和交叉验证进行超参数调试，使用 F-1 分数和 ROC/AUC 进行衡量
- 基于回归模型的系数和树的特征重要性从模型中交叉抽取重要特征，发现与同伴的比较、家庭财富和地位，以及健康最重要，并给出了相应的解释和建议；提出潜在的问题，例如样本数据的潜在偏差等

贷款发放预测/P2P 坏账建模 **2021/2019**
- **(2021)** 使用 R 语言 ggplot2 包进行探索性数据分析 (EDA) 和数据可视化，使用 SMOTE-NC 算法来平衡数据集，使用 FAMD 算法对连续和类别混合型数据生成主成分
- 使用 L-1 正则化的 logistic 回归来进行特征选择；与随机森林、XGBoost 和 ANN 进行了一个简单的比较；正确率从 65%提升到 72%
- 对现实的征信体系进行了研究，尝试确认正确率难以提高的原因，认为缺失了关键变量，例如信用记录等
- **(2019)** 在 R 中相似的预处理与建模；发现逐步 logistic 回归挑选的特征在神经网络中表现更好

高房价是否降低生育率 **2021**
- 从世界银行等来源收集数据并连接，生成面板数据；在 R 中进行数据预处理，例如填充缺失值和数据标准化；使用数据可视化来确认数据的完整性和建立对于数据的直观理解
- 使用固定效应回归来消除时不变 (time-invariant) 的混淆因子的影响；进行文献研究并确定时变 (time-varying) 的混淆因子，例如家庭财富等
- 通过对数据进行分组，考查了估计参数的不确定性. 发现高房价没有导致低生育率，甚至在低收入地区和对于富裕家庭来说有正向的作用. 给出了相应的解释