宋 骁

期望职位：数据分析

邮箱：xsong@stu.ecnu.edu.cn ◇ 18916836773 ◇ github

教育背景

华东师范大学 社会发展学院, 社会学 (均绩：3.57 ｜前 10%) *2016 年-2020 年*

相关课程：数据库基础、R 语言数据分析、算法与程序设计基础 (Python)、SPSS 应用

工作经历

泛为科技　数据分析师 (全职) *2020 年 9 月-现在*

- 使用 Apache Doris/TiDB/Clickhouse 等分布式 OLAP 数据库编写 BI 数据报表。
- 对公司的 SaaS 产品泛为智投进行用户留存分析，协助运营进行用户服务等工作。
- 对产品使用过程中的报错日志进行分析，协助产品经理优化产品设计和用户体验。
- 参与 SaaS 产品线上 BI 数据分析报表的研发工作

艾瑞咨询　数据分析实习生 *2019 年 7 月-2019 年 10 月*

- 使用 R、SPSS 软件对汽车用户进行用户画像分析工作。使用主成分分析、聚类分析法对线下数据进行无监督学习，研究不同车系用户态度区分。对汽车生产商决策提供重要参考意见，最终使用 Python 将结果进行数据可视化描述。使用 Hive SQL 访问数据库集群并协助进行数据分析。

项目经历

Kaggle M5 沃尔玛销量时间序列预测竞赛 *2020 年 6 月-2020 年 7 月*

- 使用基于 Python 的 `pandas`、`sklearn` 对沃尔玛超市分布在三个州的 10 家门店共 42840 条数据进行数据探索、特征工程分析。使用 LightGBM 算法对时间序列数据进行拟合并预测测试集 28 天的销量。最终取得前 2%(103/5558) 的银牌成绩。

中南财经政法大学　法律文本分类 *2020 年 2 月-2020 年 8 月*

- 使用 Python 的 `pandas`、`numpy` 和 `scikit-learn` 库进行文本分词、去除停用词、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。通过反向翻译方法进行 data-augment，解决训练数据不足的问题。

所获奖项

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*

核心技能

数据处理	熟悉 MySQL 数据库语言。能够快速高效地编写 SQL 代码，具有 SQL 性能优化的丰富经验 通过命令行方式操作数据库，编写 Shell 脚本进行 ETL 操作 熟悉 Python 和 R 语言对表格数据的操纵，能够使用 pandas/data.table 等工具进行数据清洗 了解统计分析，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 Tableau，R 语言 ggplot2，Python matplotlib、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
其他技能	Shell 脚本/Linux, Git, HTML/CSS/Javacript, SPSS, Stata, LaTeX, MS Office
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)