侯晓婷

- 手机： +86 189-3099-4563
- E-mail： katherine_hou@foxmail.com
- 现居地：上海市徐汇区
- 博客地址：https://monkey0105.github.io/

求职意向

高级数据分析师

工作技能

数据分析技能

- 熟练掌握Python编程, 熟练使用Python主流的数据处理，数据可视化，数据建模库，如pandas，sqlalchemy，matplotlib, seaborn, bokeh，scikit-learn等。具备能力完成从数据库提取数据，清洗数据，数据分析，数据可视化，数据建模，自动报告生成及发送的全流程。也具备一定的系统开发能力。
- 熟练掌握R语言语法，熟练使用dplyr、ggplot2包，熟练应用Split-Apply-Combine方法进行数据的转换和清洗加工。
- 熟悉Oracle, MySQL, PostgreSQL和MongoDB等数据库的数据结构，能够使用Python或R进行交互。
- 熟悉SAS语法，通过认证考试获得证书Certified Base & Advanced Programmer for SAS 9

开发工具和环境

- Ubuntu开发环境：由于Ubuntu对python更好的支持，平时的工作都在Ubuntu环境下进行，熟悉基础的shell命令。
- Git/Gitlab/Github：对代码进行备份和版本管理，并且更高效地进行团队合作的开发工作。
- StarUML2：用于绘制流程图和类图，辅助程序和系统的设计。
- Jupyter Notebook：用于探索性阶段代码的编写，或用于生成包含代码、样例数据、图表和说明文字的报告。
- PyCharm：探索性阶段完成后利用IDE的功能更好地进行debug，项目重构和维护。

语言能力

- 英语：CET4 620/710，CET6 611/710；听说读写熟练, 能无障碍地搜索和阅读英文数据技术的资料和文档。

教育背景

复旦大学 管理学院 统计学专业 统招本科毕业&理学学士学位 ***09/2009 - 07/2013***

主要课程：微积分，线性代数，概率论，数理统计，线性回归，时间序列等

华东师范大学第二附属中学 优秀毕业生 ***09/2006 - 07/2009***

工作经历

上海即富信息技术有限公司 - *金融数据部 - 数据分析工程师* ***05/2016*** *至今*

根据业务部门需要，完成ETL处理，数据分析，可视化展示，机器学习建模等工作。
根据技术部门需要，参与决策模型上线部署所需的读取模块和决策模块的开发工作。
根据数据团队内部需要，参与提高团队工作效率的Python数据处理相关功能库的开发。

上海蓝瀚广告有限公司 (WPP集团) - *MSU - 数据统计专员* ***09/2014 - 04/2016***

根据客户的需求，运用统计学知识对数据进行加工处理和建模，为客户提供专业的数据分析报告。

中国衣恋集团 - *战略企划部 - 管理培训生* ***07/2013 - 08/2014***

管理培训生组成项目组为集团内部服装、餐饮、娱乐、地产行业多品牌商业问题进行市场调查，并整理出商业分析报告，向公司和品牌领导进行汇报。

项目经历

ETL系统开发 ***10/2016 - 01/2017***

- 项目背景：
  团队在日常准备数据上消耗了大量的时间和精力。初期团队的ETL框架缺乏足够的容错性，导致ETL流程在执行时经常由于一些意外原因中断，且难以定位原因。
- 解决方案与项目成果：
  - 我完善了ETL框架，添加了遇到异常错误时重试直至最大次数的机制，并且添加了大量的日志方便错误的定位。
  - 我梳理了来自业务和建模的需求，设计、实现并测试了支付数据的ETL流程，将常用的数据定时提前计算好存入数据库用于后续的分析和建模。

机器学习风控模型的开发 ***02/2017 - 04/2017***

- 项目背景：
  公司小额贷款业务积累了一定的订单数据和坏账情况数据，同时，人工审核的效率非常低下，公司考虑上线机器学习模型进行自动化风控和审件，需要先进行线下的开发和测试。
- 解决方案：
  我整合了来自支付，小贷，同盾等多方的数据，利用随机森林模型进行机器学习建模预测小贷的坏账情况。
- 项目成果：
  模型在测试数据的表现比之前的专家模型能提升至少20%的收益。

团队基础功能库开发 ***07/2016 - 04/2017***

- 项目背景：
  团队平时在数据的读取，处理，分析，建模，可视化，报告生成等各环节存在很多重复性的步骤和工作。
- 解决方案：
  我参与了团队内部通用基础库的开发，针对重复出现的工作和需求开发相应的Python工具。
- 项目成果：
  减少团队在重复性工作上消耗的时间和精力，大幅提高了团队的工作效率和产出。

通话详单数据清洗 ***05/2017***

- 项目背景：
  将10000多个格式不统一，甚至不是标准表格形式的通话单Excel文件清洗处理成标准格式，方便后续的分析进行。
- 存在的困难：
  - 清洗的流程非常复杂且涉及大量的条件判断，很难用常规的if..else..条件判断去描述这种复杂的工作流。
  - 需要不断的通过人工观察，修改清洗处理流，并记录不能被正确处理的文件，反复迭代。
- 解决方案：
  我专门开发了相应Python库用于构建和可视化带有多个条件判断的复杂处理流，以及交互式的升级处理流并且维护清洗结果的工具，以辅助这个工作的顺利进行。
- 项目成果：

开发的Python库可供今后类似场景长期使用；清洗出了70%的文件数据，并进行了初步分析和可视化展示。

运营数据自动邮件报告 ***03/2017***

- 根据业务部门需求，设计和计算运营指标数据，并且按照日，周，月的频率发送邮件报告。
- 从数据库提取数据，统计计算，可视化做图，HTML报告生成，邮件自动发送全环节做到全自动化程序化。

流失用户分析 ***08/2016 - 09/2016***

- 分析每个用户的交易历史，绘制用户流失曲线。由于数据量庞大，无法利用pandas，主要运算通过SQL完成。
- 对流失用户和留存用户特征进行对比分析，找出流失用户的典型画像，对即将流失的用户提出干预方案。

非标准化错误信息文本归类 ***09/2016***

- 利用jieba分词模块进行文本分词和分类，将数据库中300多种非标准化的错误信息归纳成20类，方便后续分析的进行。

展厅访问量模型/网络点击量模型 ***09/2014 - 12/2014***

- 项目背景：
  客户在不同的媒体渠道不同程度地进行了广告投入，进而想了解这些投放对于销量/展厅访问量/网络点击量的影响，并且希望借鉴经验来对接下来的预算进行调整。
- 解决方案：
  建立展厅访问量/网络点击量与不同媒体渠道投放金额间的关系模型，测算不同渠道的广告效率。以最大化展厅访问量/网络点击量为目标，在有限的投资预算下，为客户提供最优媒体投放计划。
- 存在的困难：
  数据维度大于数据量，媒体变量的非线性变换参数选择。
- 解决方案与项目成果：
  我和经理两人合作设计和开发了R交互式建模工具StepReg，将模型拟合和参数调整的思路融入程序设计和编写中，提高每次更新模型的效率至两天内。

文本分析工具开发 ***06/2015***

- 对问卷收集/网络论坛上获得的英文文本进行分词，进行态度分析和主题分析，绘制词云图和词关联图。

培训经历

业余时间在各MOOC平台对大数据和机器学习领域进行学习。

- Deep Learning Nanodegree Foundation, Udacity *进行中 03/2017*
- Introduction to Computer Science & Programming Using Python, MIT, edx *Achieved 10/2015*
- Core Presentation Skills Workshop, Simitri *Achieved 06/2015*
- R Programming, Johns Hopkins University, Coursera *Achieved with Distinction 11/2014*

个人荣誉

- 复旦大学 毕业生奖学金 *二等奖* *05/2013*
- 上海世博会 优秀志愿者团队 *复旦大学世博白莲泾出入口小组* *11/2010*

感谢您付出宝贵时间阅读我的简历！