夏强

- Github: https://github.com/ShawnXiha
- Email: huakaijianwo2009@163.com
- Phone: 18679188050
- Location: 江西南昌

- **摘要**

数据分析师，能熟练使用Python及R数据可视化数据分析数据清洗，熟悉并会使用常用的分类和回归算法对数据进行分析预测。

- **技能:**

- 语言： Python,R,Matlab,Javascript,HTML5,CSS3
- 库: NumPy,Pandas,Matplotlib,scikit-learn,d3.js,dimple.js,ggplot2,tensorflow,genism,keras,xgboost
- 其它: git/github,SQL,MongoDB

Udacity项目

1. **调查数据集**

- 使用Python Pandas库整理清洗数据集，使用Matplotlib数据可视化分析数据分布
- 使用Python scipy库对各因素对结果影响进行假设检验

2. **探索并总结数据**

- 使用R ggplot2库数据可视化分析数据分布，特征之间关系
- 使用R GGally库探索变量之间的相关性

3. **从安然的邮件中识别舞弊**

- 发现并移除数据集中的离群值，异常值，使用主成分分析及特征选择对机器学习的数据进行预处理
- 选取性能度量指标，采用交叉验证选择最优的机器学习算法并调节参数

4. 处理 OpenStreetMap 数据

- 应用python解析上海市开放地图xml数据并发现其中存在的异常
- 应用正则表达式修改地图数据中英文混合，中文及拼音混合，英文的街道名修改成中文名
- 将数据导入MongoDB，并用MongoDB分析探索地图的基本信息

5. 高效数据可视化

- 发现数据规律，并使用JavaScript d3.js可视化分享发现
- 接收反馈，并根据反馈意见提高数据可视化的效果

6. A/B 测试

- 使用A/B测试分析，Udacity如果在注册学习前提示每周最少需要投入的学习时间能否可以在减少免费试学学生数量的同时不减少付费学生数量，并判断这个决定可不可取。
- 设计一个后续试验来减少免费试学后取消付费学生的人数。

7. 监督督学习：构建学生干预系统

- 根据各种监督学习算法的优缺点及当前数据集的特点，选择三种适用的算法；
- 综合考虑各个算法在数据集上的表现如训练时间，测试时间，在测试集上预测的精度，选择最佳的算法；
- 使用交叉验证调节算法的超参数；

7. 非监督学习：创建用户细分

- 使用主成分分析对数据进行特征转换；
- 使用高斯混合模型聚类算法对顾客购买特征创建聚类

8. 强化学习：训练智能出租车学会驾驶

- 使用 Q-learning算法，训练人工智能体，使它能够对周围环境做出最佳选择。

9. 拍卖中的机器人识别

- 特征提取：从七百六十万次拍卖出价数据中提取每个拍卖者的特征
- 特征工程：探索分析数据并对其数据变换，缺失值填充，标准化处理
- 模型选择与优化：应用优化随机森林，极端随机树，梯度提升树算法对拍卖者进行分类。最后对测试数据的预测结果roc_auc值为0.909

kaggle比赛项目

1. **高级回归技术的应用：房价预测**

- 特征工程：分类特征one-hot编码，对分布倾斜特征对数运算，缺失值填充
- 模型选择与优化：应用岭回归和套索回归对房价进行预测。采用两种算法对房间预测的平均值作为预测结果，均方误差为0.11953。

2. **Two Sigma Connect: Rental Listing Inquiries**

- 特征提取：应用Word2Vec词向量模型将网上租房信息中的房间描述文本转化为特征向量
- 应用梯度提升树对顾客对房间感兴趣程度是低，中，高的概率进行预测。最后对测试集预测的logloss值为 0.57193

教育

在线教育

1. **Udacity** 数据分析师纳米学位 2016.6~2016.9
2. **Udacity** 机器学习工程师纳米学位 2017.1~2017.3
3. **Coursera** 机器学习by吴NG 2016.10~2017.1

学历

1. **南昌大学** 食品科学硕士 2013.9~2016.6
2. **江西农业大学** 轻化工程学士 2009.9~2013.6

其他技能:

设计、计划、安排、实施实验并分析处理数据写成文章

做ppt，演讲