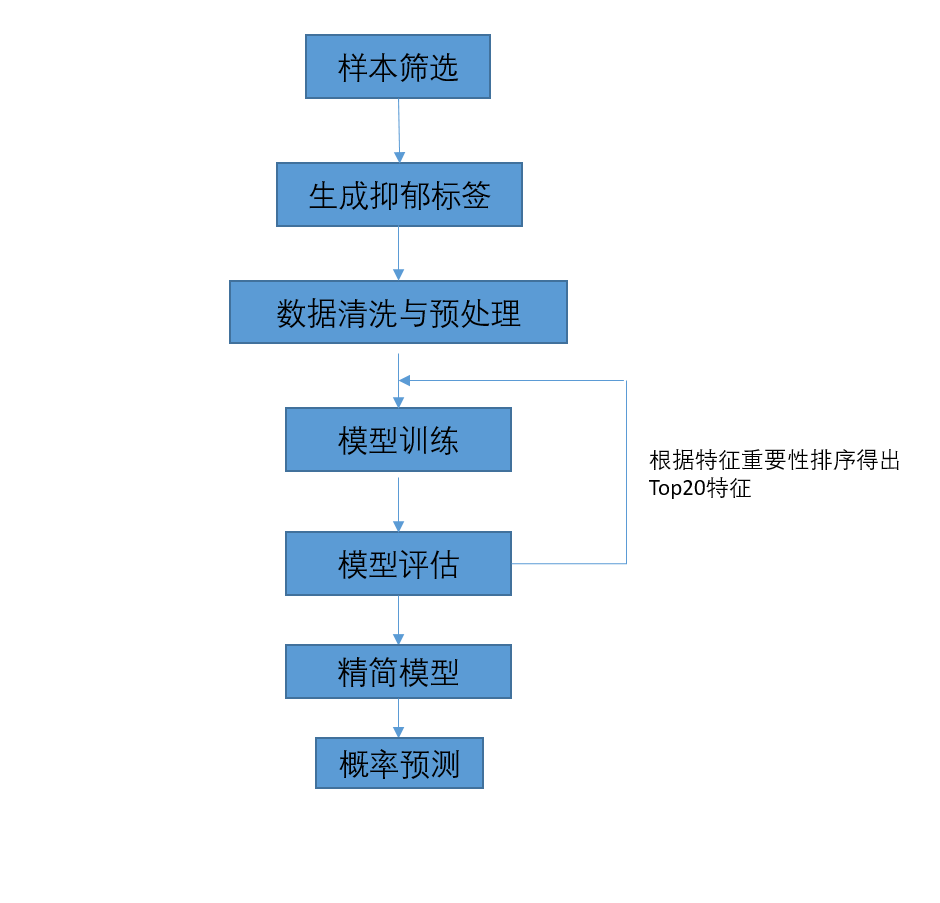
1. 流程概述

本项目旨在对60岁及以上的中国社区老人是否抑郁进行预测。目前现有的全年龄样本量有20547人，清理之后60岁及以上的样本有一万多个。我们基于这一万多个的样本构建数据集，同时根据CES-D表的得分来区分样本抑郁与否（总分30分，大于等于12分为抑郁，标签为1，小于12分为非抑郁，标签为0）。构建模型前，我们会将基线（2013）已经发生抑郁状况的老人进行剔除。

我们对数据集分别训练不同的机器学习分类模型：逻辑回归、K近邻、朴素贝叶斯、决策树、支持向量机、随机森林、梯度提升决策树、Xgboost、Catboost共9个模型，并且在训练结束后通过计算模型的曲线下面积（Area Under the Curve [AUC]）、准确率（Accuracy）、精确率（Precision）、灵敏度(Sensitivity)和特异性(Specificity)指标，对模型表现进行多维度评估。在此基础上，我们将从中选出综合表现最好的5个模型，采用五折交叉的方式来构建模型融合(Model Stacking)的集成方法来训练数据。模型融合是一种通过结合多个模型来改善机器学习模型表现的集成方法，常用于许多著名的机器学习竞赛平台。模型融合的基本思想是同时训练多个算法原理不同的机器学习算法，并在此基础上训练一个元模型来对不同的模型继续有机融合，最后基于这些模型返回的多个预测结果来输出最终的预测结果。模型融合的优势在于能集成不同模型对数据集内在关系的抓取，以期达到优于单一模型表现的效果。最后，我们将模型融合算法与选出的5个表现最好的模型放在一起再做一轮评估来找出最优模型。用选出的模型对2015年的样本数据进行预测两年内的抑郁发病情况，得到每一位老人发生抑郁的风险预测概率（百分比）。

另外，再我们训练基于树模型(tree-based model)的机器学习算法时（诸如随机森林、梯度提升决策树等），通过计算基尼重要性来排列出具有较高重要性的特征变量。我们将选取其中重要性最高的20个变量，并将这20个变量从原始数据集中筛选出来重新代入我们选出的最优模型框架，并将训练结果与原模型进行比较，在曲线下面积、准确率、精确率、灵敏度和特异性指标相差不大的前提下得出最优的特征选择组合。以此来指导我们在设计问卷过程中需要注意的重要问题。



数据库介绍

中国健康与养⽼追踪调查是由北京⼤学国家发展研究院 主持、北京⼤学中国社会科学调查中⼼执⾏的⼤型⻓期追踪调查项⽬。中国健康与养⽼追踪调查⾸先在⽢肃、浙江两省进⾏了包括 1,570 ⼾家庭中 2,685 个 45 岁以上中⽼年⼈的基线预调查，并且在 2012 年成功地进⾏了追踪访问。在此基础上，中国健康与养⽼追踪调查于 2011-12 年在全国进⾏了⼤规模基线调查。这是 具有全国代表性的随机抽样调查，覆盖了不包括西藏在内的中国⼤陆所有县级单位。 样本包括了 150 个县级单位（散布在全国 28 个省区，如图1.1所⽰），450 个村级单 位，10,257 ⼾适龄家⼾中的⼀个⾄少年满 45 岁的⼈，包括其配偶，共 17,708 ⼈。2013 年，项⽬组圆满完成了全国基线样本常规调查的第⼀次追踪访问，追访成功率达到 88%。2014 年夏天，CHARLS 对全体受访者进⾏⼀次⽣命历程的专 题调查，对 12,250 样本⼾中的 20,547 位受访者⾃出⽣以来在家庭、婚姻、⽣育、就 业、教育、迁移、健康等⽅⾯的历史全景进⾏描述，追访成功率为 86%。我们将使用2013年的调查数据和2014年的生命历程调查数据来构建抑郁风险预测模型。

2013年共纳入变量6440个，分为以下13个类别：



2014年共纳入变量6955个，分为以下7个类别：



1. 数据清洗与预处理

具体包括：

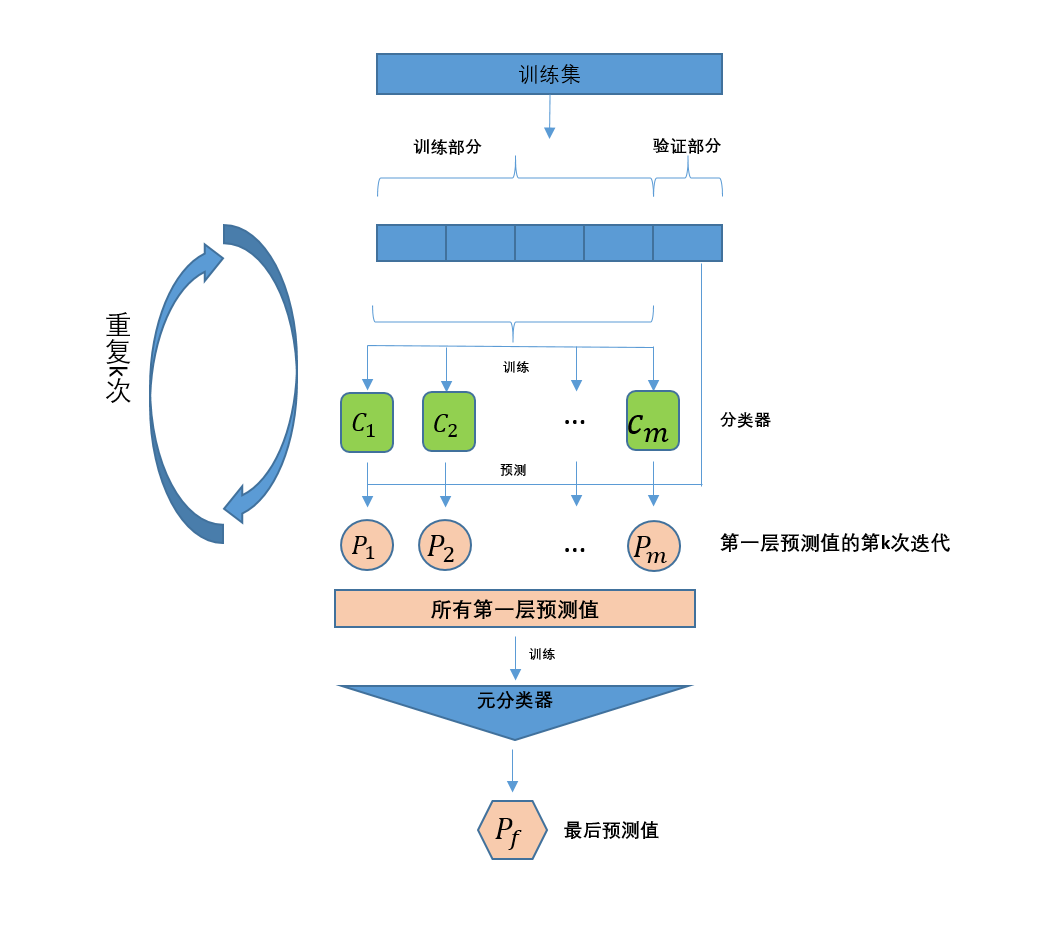
* 去除重复样本
* 去除缺失率较高的特征变量（通常认为超过70%的数据缺失率不可用）
* 特征变量的逻辑筛选（如身高5米或者体重10千克）
* 独立热编码（one-hot encoding）,用于处理类别变量
* 数据标准化，无量纲化且加快算法收敛速度

其中在缺失值填充阶段，我们考虑采用K-近邻算法（K Nearest Neighbor）进行缺失值填充，它通过计算 高维空间上样本之间的距离来识别相邻点，并利用相邻观测值的完整值来估计缺失值[1]。KNN的优点有：a)插补值是数据集中实际出现过的值，并非是二次计算得来的。b)根据变量信息保留了原始数据的分布结构。c)KNN插补完全是非参数的[2]。这就意味着这个基于KNN算法的新方法使得我们可以更便捷地处理缺失值，并且与直接用均值、中位数相比更为可靠。

1. 类别不平衡问题的处理

对于在实际建模过程中存在的目标变量类别不平衡问题（既样本中预计发生抑郁人数比例较高，导致两个类别比例极不均衡），我们拟采用合成少数类过采样技术(Synthetic Minority Over-sampling Technique)来扩充少数类样本[3]。它是基于随机过采样算法的一种改进方案，由于随机过采样采取简单复制样本的策略来增加少数类样本，这样容易产生模型过拟合的问题，即使得模型学习到的信息过于特别(Specific)而不够泛化(General)，SMOTE算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中，算法流程如下：

1. 对于少数类中每一个样本,以欧氏距离为标准计算它到少数类样本集中所有样本的距离，得到其k近邻。
2. 根据样本不平衡比例设置一个采样比例以确定采样倍率N,对于每一个少数类样本,从其k近邻中随机选择若干个样本，假设选择的近邻为。
3. 对于每一个随机选出的近邻，分别与原样本按照如下的公式构建新的样本
4. 模型融合算法原理示意图：



图片来源：https://rasbt.github.io/mlxtend/user\_guide/classifier/StackingCVClassifier/

其中分类器共有9个，分别为逻辑回归、K近邻、朴素贝叶斯、决策树、支持向量机、随机森林、梯度提升决策树、Xgboost、Catboost。

1. 模型评估指标说明

*混淆矩阵*

True Positive (真正，TP) ：将正类预测为正类数

True Negative (真负，TN) ：将负类预测为负类数

False Positive (假正，FP) ：将负类预测为正类数

False Negative (假负，FN) ：将正类预测为负类数

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| True | True Positive(TP) | True Negative(TN) |
| False | False Positive(FP) | False Negative(FN) |

1. 准确率（Accuracy）

准确率（Accuracy）计算公式为：

准确率就是被分对的样本数除以所有的样本数，通常来说，正确率越高，分类器越好。

1. 精确率（Precision）

精确率（Precision）定义为：

表示被分为正例的示例中实际为正例的比例

1. 灵敏度（Sensitivity）

表示的是所有正例中被分对的比例，衡量了分类器对正例的识别能力。

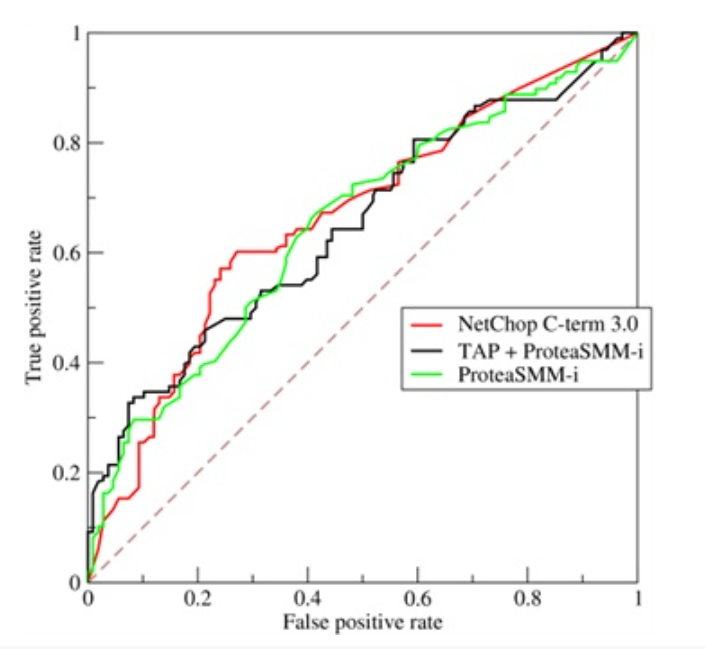
1. 特异性（Specificity）

表示的是所有负例中被分对的比例，衡量了分类器对负例的识别能力

1. 曲线下面积（Area Under the Curve[AUC]）

ROC (Receiver Operating Characteristic) 曲线一般横轴是False positive rate, 纵轴是True positive rate。而AUC既为曲线下的面积，一般AUC值越大，说明模型越好。

*曲线示例*：



1. 特征变量选取

特征变量选取基于计算特征重要性，在预测建模项目中起着非常重要作用，能够提供对数据、模型的见解。

特征重要性是指，在全部单颗树上此特征重要性的一个平均值，而单颗树上特征重要性计算的方法为：根据该特征进行分裂后平方损失的减少量的求和。

计算原理如下：

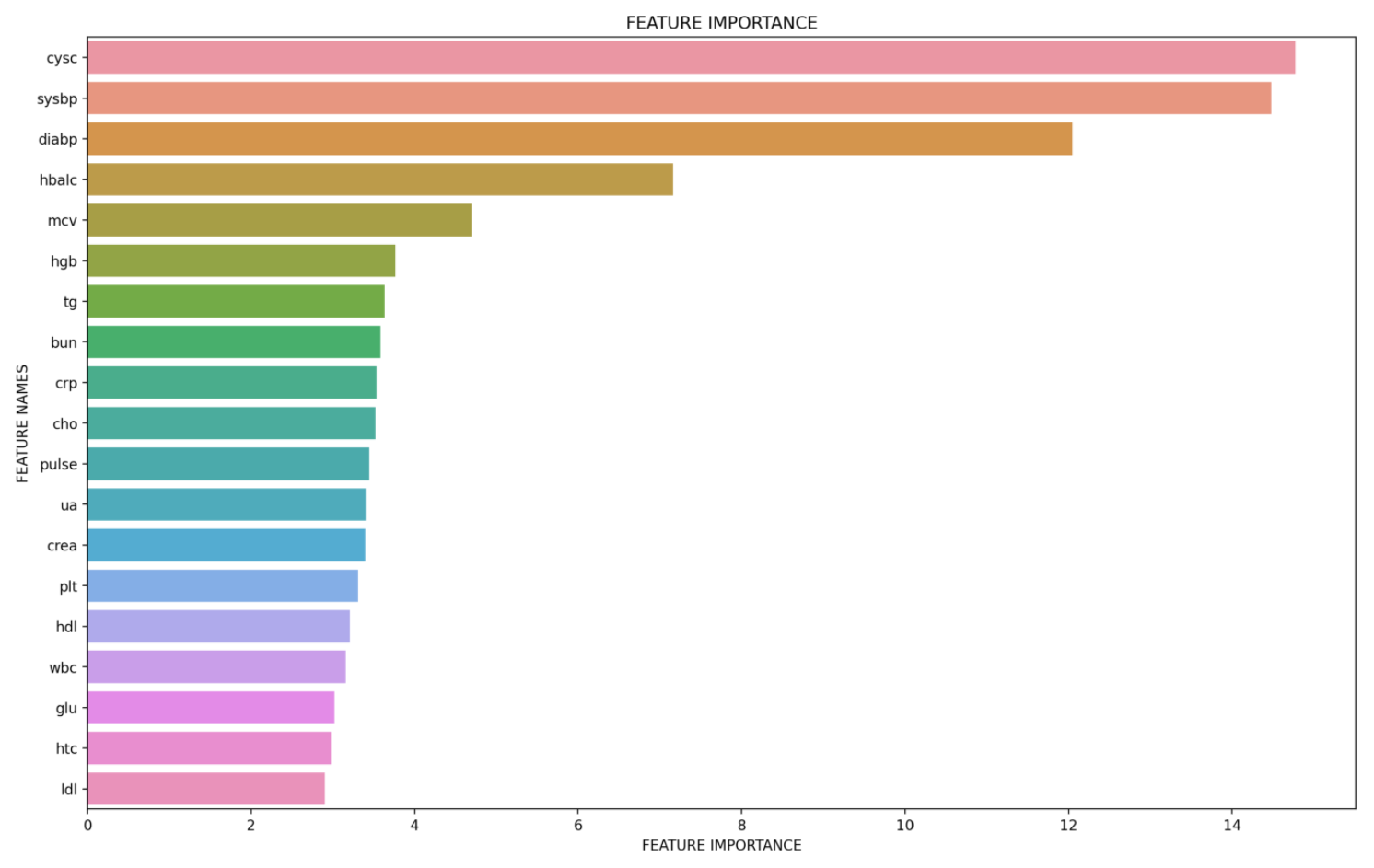
* 特征在整个模型中的重要程度为：

其中，M是模型中树的数量。

* 特征在单独一个树上的特征重要度为：

其中，是树中非叶子节点数量，表示在内部节点t进行分裂时选择的特征，是内部节点分裂后平方损失的减少量。

*特征重要性示例图：*



1. Beretta, L. and A. Santaniello, *Nearest neighbor imputation algorithms: a critical evaluation.* BMC Medical Informatics and Decision Making, 2016. **16**(3): p. 74.

2. Su, D., et al., *Use of machine learning approach to predict depression in the elderly in China: A longitudinal study.* J Affect Disord, 2020. **282**: p. 289-298.

3. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* 2002. **16**: p. 321-357.