附录1专题选题

专题一：基于“中国家庭追踪调查”（CFPS）的数据发现和挖掘

分主题 1 ：预测家庭样本的流失。参赛者在 CFPS 2016 年发布的家庭关系库（数据下载地址为：http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/45LCSO）中近 15000 个 fid16 中选出 1000 个最有可能在 2018 年流失的家庭。CFPS 2018 实地工作结束后我们根据执行的最后结果选出命中率最高的参赛作品。

分主题 2 ：收入数据的插补。由于收入数据较为敏感，在抽样调查中会出现一定比例的缺失情况。参赛者针对缺失以及可疑的收入数据提出插补方案并给出插补结果。我们将组织相关方面专家对方案的合理性以及最终结果进行评估。

以上两个主题均不限研究方法，传统的统计模型或机器学习方法均可。

附加要求：①需要提交参赛源代码至北京大学开放数据平台，代码需要为 Python、R 或其他编程类语言代码；②需要有说明文档描述代码的运行环境和使用方法；③代码结构清晰，有适当的注释。

专题二：社会经济调查的职业和行业自动编码模型构建

社会经济调查中通常会采集职业和行业信息，为方便数据用户使用这些信息，一般会事先基于国家标准化管理委员会发布的《职业分类与代码》对上述信息进行编码。组委会将在竞赛平台上提供部分社会经济调查中采集得到的职业和行业的具体描述信息，以及相应的已经编码成功的代码。要求参赛者基于上述数据（数据下载地址为：http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/PEMXPX），构建自动编码模型。组委会将利用该模型，应用于其他已人工编码成功的数据。基于模型预测的准确度，评判模型的优劣。

附加要求：①需要提交参赛源代码至北京大学开放数据平台，代码需要为 Python、R 或其他编程类语言代码；②需要有说明文档描述代码的运行环境和使用方法；③代码结构清晰，有适当的注释。