选题报告

## 问题描述

本组选题为“（2）参加一些著名的机器学习相关竞赛”。

本组所选的竞赛赛题为“【长期赛】- 测测你的一见钟情指数”，在阿里云的天池大赛平台上举行。其链接为 [【长期赛】- 测测你的一见钟情指数-天池大赛-阿里云天池 (aliyun.com)](https://tianchi.aliyun.com/competition/entrance/531825/introduction)。

该赛题的赛题简介如下：

“性格判断，职业预测，恋爱指数……你一定在各种网页或微信公众号里接触过此类的测试小游戏。只需要回答几个简单的问题，长篇大论的分析文章就被甩到你脸前，说得头头是道，看起来也有那么几分道理。在信息时代的大背景下，利用数据分析来实现人格测试早已屡见不鲜。

“本次学习赛的目的，正是利用阿里云的机器学习平台PAI DSW来实现一个相亲战斗力检测器，在教授技术的同时也带大家层层剖析此类游戏的原理。

“本次一见钟情学习赛，受到哥伦布亚商学院教授Ray Fisman和Sheena Iyengar联合发布的文章《伴侣选择中的性别差异(Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment)》启发，决定利用机器学习的数据实验来分析当下男女在相亲交友时，个人的内在外在各类因素对最终相亲结果的影响。

“在2002年-2004年期间，Ray Fisman教授和Sheena Iyengar教授在筹备论文时，邀请志愿者参加闪电速配实验(相亲车轮战，每4分钟与一名相亲对象快速沟通，然后再换下一个相亲对象)，提供一些相关的个人信息给相亲对象，并询问相亲对象给出是否愿意在不久的未来再次见面。本次学习赛的分析数据，记录了当时一见钟情相亲实验时，志愿者的相关信息及相亲结果。

“数据集的内容包括实验志愿者的性别、年龄、人种、专业、地区、收入等特征，以及志愿者对配偶是否来自同一地区、同一信仰等观点的预期。

“选手可以针对数据集不同字段间的相互影响进行分析，训练一个机器学习模型，去预测实验人身上一个或多个特性对其相亲成功与否的影响。也就是利用其它特征信息，预测数据集中的“match”字段的结果，1=成功，0=不成功。

“训练完模型后，在‘提交结果’页上传自己的模型即可。提交后将进行实时测评，每天每位选手可提交2次；排行榜每小时更新，按照评测指标得分从高到低排序；排行榜将选择历史最优成绩进行展示。”

## 相关文献调研

本题的数据集来自于哥伦布亚商学院教授Ray Fisman和Sheena Iyengar筹备文章《伴侣选择中的性别差异(Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment)》时邀请志愿者进行一见钟情相亲实验所得到的实验数据。

不过，该文章更多地是试图通过数据分析性别差异及其原因，与我们机器学习的目标没有直接关系，但可以根据文章中的内容来考虑各个属性的权重，提高模型的准确度。

## 初步拟定的解决方法

本赛题所提供的数据集是csv表格，存有当时相亲实验的志愿者的各种信息。这些信息有的是数字，有的是字符串，而且有大量字段缺失。因此需要首先对数据集进行一些处理，使其适用于机器学习。

最基本地来看，该问题是一个二分类问题。因此可以考虑SVM，Logistics回归，决策树等模型。决策树模型的选取也有很多种，比如4.5，CART，随机森林，boosting等。而且因为该问题数据有很多是标签类离散数据（比如职业等），优先考虑采用决策树类算法。

本组计划各小组成员对多种方法都进行实验，根据实验结果，对效果较好的方法进行进一步优化、改良和调参，争取实现一个效果较好的模型，在竞赛中取得好的成绩。

## 工作计划

第 11-12 周：小组成员尝试并实现多种机器学习的模型，以各种决策树算法为主，如果时间足够考虑尝试SVM，逻辑回归等。

第 13-15 周：小组成员选取效果较好的模型，进行进一步优化、改良和调参，争取实现一个效果较好的模型。

## 组员姓名学号信息

组长：沈冠霖 2017013569

组员：符景洲 2018013419 段凡 2018013398 杨天煜 2018013364