基于机器学习算法的网络信贷风险评估

注：本文源自笔者2020年本科毕业论文，已收录在兰州大学图书馆。本文旨在分享学习，请勿作他用。

项目结论

本文基于前人的研究成果，通过人人贷网贷平台的数据，对借款人是否会违约进行了预测，先后使用了LDA、决策树与 SVM 三种算法训练预测模型，并针对 LDA 算法进行了尝试性改进，经过各项指标权衡，最终选择改进后的 LDA 算法作为最优模型.

在训练模型的过程中，本文得到了如下结论：

1. 经过多个模型拟合数据、对比结果，从精确度，召回率，F1值，准确率各方面考量，结合网贷实际情况，本文认为改进的 LDA 模型更加适合处理贷款违约预测问题.
2. 根据决策树及 SVM 训练结果，我们发现还款期限、年利率、还清笔数、信用额度等特征对违约预测有相当程度的影响，特别是期限三年以内，年利率较高的贷款较容易发生违约. 与之相反，性别、学历和房贷的影响很小. 相关人员可以加以关注、借鉴.
3. SVM 整体性能最佳，对于更加重视模型总体指标的机构或者借贷交易量较少的小型网贷机构可以优先考虑选择线性 SVM 进行相应业务预测.

本文各个模型取得较高精度的原因：

1. 在数获取的过程中，本文使用了爬虫从平台官网采集样本，官方公布的数据经过了平台方面的必要处理，使得数据更加规整.
2. 样本多数特征在经过初步筛选后，本身便与个人财务有着直接相关性，与客户是否借贷、会否违约有较强关联.
3. 训练模型前对数据进行了标准化，使得各个特征的权重相对均衡，不会因为某些特征在初始数据样本中比重更大而影响了其他特征的分类作用.

纵观整个研究过程，本文也存在以下不足：

1. 采集的样本数量较少，噪声对模型有一定程度的影响，训练过程中可能将噪声的特点也学入了模型，影响最终结果.
2. 虽然使用十折交叉验证尽可能的减少了模型的偶然性与意外性，但对于是否存在过拟合问题，本文未进行深入探查.
3. 在使用 LDA 与 SVM 相关算法时，需要各个特征之间是相互独立的，而本文在默认样本数据满足这一条件的基础上进行了相关训练，虽然最终结果良好，但过程并不严谨. 文中没有进行特征非线性相关性的检测，我们无法断言散标样本的特征之间完全独立.
4. 在决策树算法中，每一次交叉验证得到的树结构都不完全相同，即树不够稳定，可以使用随机森林或其他集成方法训练出更加稳定、准确的模型.

对于本文使用的研究方法，事实上还有一些可以进一步提高模型精度的方法，比如利用决策树得到的 6 个划分变量结合 SVM 或 LDA 进行更细致的预测，但由于现有模型精度已足够，过于追求更高的性能看起来锦上添花，但耗费了资源却没有取得明显效果，实则画蛇添足，也有悖奥卡姆剃刀原理.

随着科学的蓬勃发展，机器学习算法也在不断创新、完善，但不论如何变化，机器学习的初衷是不变的——利用机器强大的运算能力，结合恰当的计算方法更好地解决实际问题，造福人类社会. 基于这一理念，我们应具体问题具体分析，因地制宜，因时而定，不同的场合使用不同的算法，正确灵活地运用算法解决实际问题，才能让机器学习在人们生活中大放异彩.