我们小组研究的课题是基于 stacking 算法的模型融合来构建 P2P 网贷中信贷用户违约预测模型。

本次介绍将从：选题背景、文献综述、研究内容和条件保障这四部分展开

首先是选题背景。

作为一种依靠于互联网的金融业务平台，P2P 网贷能满足小微型企业和个人借款者的需求，又能为投资者提供较高利率，在我国的发展十分迅速，就像这张图中所展现的，P2P 网贷规模逐年上升。但与此同时我们也应该看到，P2P 平台本身就存在着巨大的信用风险，在目前各平台信用评估体系不健全的状况下这一风险更加严重。因此，我们也可以发现，在 P2P 平台一路高歌猛进的同时伴随着一系列的倒闭狂潮。

近几年，在同业竞争的压力下，许多 P2P 平台偏离传统模式，为吸引客户开展了一系列新的业务活动，如图像上展示的债权转让模式、担保模式和平台模式等。这些新业务开展的过程中，P2P 平台开始涉足线下活动、担保活动并直接参与了借贷双方的资金流动过程，但这些平台并不具有被国家承认的开展此类业务的资质，并且在开展这些业务时也时常出现平台拿自有资金去偿还坏账的情况，这更加剧了平台面对的风险。

在 P2P 网贷平台的风险不断暴露的同时，国家开始重视对 P2P 平台的监管，对平台的管理模式、经营模式和信息披露模式等提出了一系列的整改措施，这也意味着 P2P 平台通过开展新的业务模式来提升竞争力已经行不通了，此时最有前景的出路便是在纯平台模式下提升其自身经营管理能力，建立健全其风控系统，这样才能保证其自身的平稳发展，这也是我们课题小组的着眼点。

在文献综述方面我们选取了部分具有代表性的文献来展现我们课题的推进过程。通过文献 1，我们能知道目前我国存在哪几种 P2P 平台的模式以及各种模式存在的共性和特性的风险。文献 3 提出互联网信贷的高风险来源于缺少征信管理，文献 2 论证了信用评级体系确实对平台发展有利。由此，我们可以看到，建立信贷用户违约预测模型对 P2P 平台的发展确实是有意义的。

风险预测模型刚开始发展时多为单一模型。单一模型有各自的优点，但是也都有不完善的地方，不能全面地进行风险评估，在此基础上便出现了组合类的优化模型，并且文献 12 从信用风险预测准确性角度探索了五种不同类型噪声分类器预测行为，为集成模型比单一模型有更高的预测准确率这一观点提供了理论依据。

目前存在的组合模型主要有两种：

第一种方法就是把单一模型重复运用综合考虑而得到的“合弱成强”式模型。代表方法

为 bagging（如随机森林）、boosting（如梯度提升树）。文献 9 以决策树作为基本算法，采取集成机器学习来进行信用评估，这使它的结果的整体正确率从其他方法的 83.25%增加到 85.86%。

第二种便是把多个不同类型的单一模型组合而成的“强强联手”式模型。文献 10 就将 stacking 方法与 bagging、boosting 模型结合用于风险评估。

现在我们知道，组合类的模型要优于单一模型，而目前存在的这两类组合模型又各有利弊，没有高下之分，故我们小组结合我们的问题的特点选用运用 stacking 方法的组合模型来构建我们的模型。

第三部分是我们本次研究的主要内容，我们的研究目的是降低信贷用户违约的风险，我们要做的就是建立一个模型来实现我们的目的。在研究的过程中我们要考虑是我们要建立的模型受那些因素的影响？我们选用哪些模型来拟合更合适？还有模型建成后又如何对我们的模型进行评估?

根据我们研究思路的指引，我们可以得出我们研究的主要内容。首先，我们要根据银行等实体金融机构的经验和前人的研究，在结合 P2P 平台自身特点的基础上完善我们的指标体系；然后，我们选用几种适用于此问题的单一模型，并用 stacking 算法进行单一模型的融合来构建预测模型。再次，我们用 AUC 等方法来进行模型的评价，并进一步发现问题，优化模型。最后，利用我们所构建的模型来对 P2P 平台的发展提出可行的建议。

我们用来实现以上研究内容的方法主要有：文献研究法、比较分析法、调查研究法、个案研究法和模型构建法。值得一提的是，我们小组同时运用了探索性研究的方法，在小范围内进行数据的处理和模型的构建，以验证我们所构建的模型的可行性和正确性。下面是对我们的探索性研究方法的一个简单的介绍。

我们选取了一个台湾省信贷客户违约情况的数据集。Y 是响应变量，表示是否违约，1 表示违约，0 表示未违约，其他变量为解释变量，包含可能对用户违约行为产生影响的的一系列指标。

由于原始数据存在共线性、类别不平衡等诸多问题，拿到数据之后还需对数据做进一步的加工才能使用，例如要进行离散化以增强模型的鲁棒性，标准化以消除量纲的影响，变量降维以消除变量相关性影响

从变量降维这一步来说，这里展示的是一张变量间的相关矩阵图，可以看出连续变量 X12-X17 存在较大相关性，这几个变量分别表示客户 4 月到 9 月的信用卡消费情况，为了剔除这几个变量中存在的冗余信息，降低建模的复杂度，我们采用主成分分析方法对这几个解释

变量进行降维处理。主成分分析通过对协方差矩阵做奇异值分解，将大量相关性很高的变量转化成几个相互独立、且能解释大部分原始数据信息的主成分。（如果能写在 ppt 上就不用我解释了）主成分分析的结果如表所示，从表中可以看到第一主成分就贡献了原始数据 90%以上的信息，因此我们用第一主成分代替 X12-X17 这六个解释变量。查看第一主成分与这六个变量的系数可以获知其实际意义为“月均消费金额”。

对数据做完初步处理之后，我们利用分层随机抽样按照 7:3 的比例将原始数据划分成训练集和测试集两部分。模型的建立过程中，对于所有的基学习器，均采用 5 折交叉验证的网格搜索法来确定重要超参数的取值。选择分类正确率和 AUC 值作为模型的评价指标。分类正确率是正确分类的样本数占总样本数的比例，反映了模型的整体分类精度。AUC 值代表模型 ROC 曲线下的面积，常被用来判断一个二值分类器的优劣，是对正例分类精度和反例分类精度的综合度量，AUC 取值越接近 1 表示分类器的效果越好。

模型的预测结果如表所示。由图表我们可以看到，Stacking 集成模型的分类正确率最高达到了 xxx%，比最好的单一分类器 xxx 高了 xxx%，这说明 stacking 集成能在一定程度上提升分类效果。AUC 值的结果同样体现了 Stacking 集成模型有着最优的泛化性能

在对模型对现实的指导这一点上，除了进行违约预测外，通过模型输出结果我们可以得到各个变量对模型的贡献程度，可以看到哪个变量更重要，对客户违约率影响较大，平台日后可着重收集类似的信息，贷款方也可以加大对这方面的重视。

第四部分便是我们小组的条件保障，首先，小组成员均来自统数学院，接受过良好的数学与统计教育，数理功底较好。而且有的同学参加过专业的数据分析培训，参加过许多比赛，这些都将有利于小组开展研究。

在调研渠道方面，小组主要通过文献和电子资源获取相关资料，并且我们就相关问题咨询了光大银行成都分行行长周楠先生，受到科研经验丰富的潘蕊老师的悉心指导。

小组预计申项成功后，6 月份左右开始获取人人贷数据，10 月份左右完成中期报告，明年 3、4 月份能够顺利结项。并且每一阶段的成果都会由文档等形式予以展现。