企业经营退出风险预测 解决方案

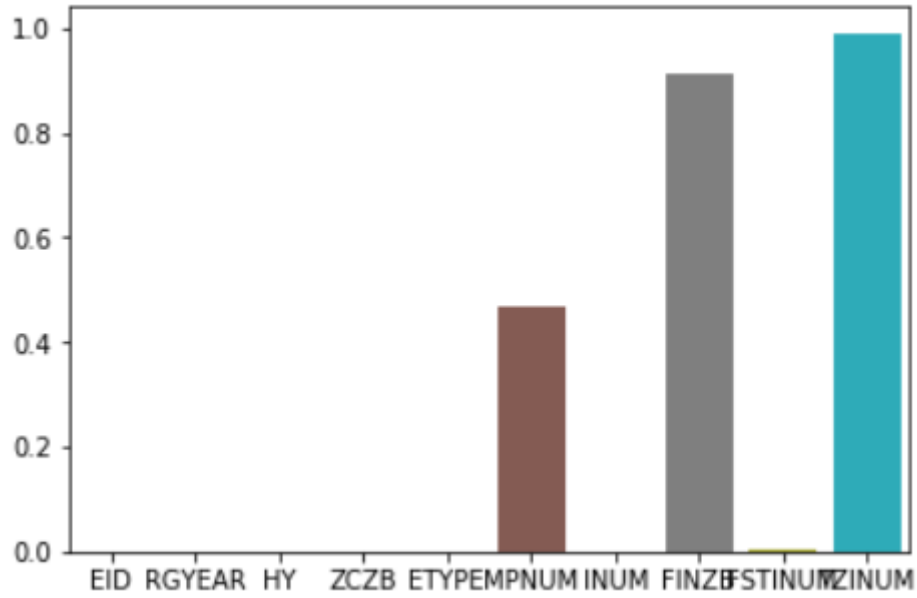
一、解决方案概述

企业经营退出风险预测的数据总计有15W条已知企业经营状况的训练样本，以及企业其他业身份信息（已脱敏）及企业在一定时间范围内的行为数据。因此需要我们从这些行为数据中提取企业相关数据来构建训练集。数据存在类别不平衡，特征构建难度高等特点。本次方案用途主要是为了能从企业的基本信息数据、变更信息数据、分支机构数据、投资数据、权利数据、项目数据、被执行数据、失信数据以及招聘信息数据这一系列公开、透明、可收集的系列数据中挖掘出企业倒闭的概率，从而为投资者做一个风险预测。
针对需要解决的问题和数据特征，我们主要从以下几个方面来进行处理：数据预处理、特征组合、特征工程以及算法选择。

二、从原始数据中组合训练特征

由于数据提供方提供的训练数据仅有企业ID与标签，因此本次解决方案的第一个难点就在于如何科学的生成训练集。因此需要我们才数据提供方所提供的企业行为文件中提取出可能有用的信息来作为建模的特征，正如那么一句话在业界广泛流传：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。因此本文接下来会着重介绍特征选取的思路。

1.企业基本注册信息数据
企业基本数据文件总计有25W多条数据，基本包括了训练集以及测试集的所有数据，因此直接将该文件按照企业id组合。
如下图所示，其中这四个字段有一定的缺失数据，根据数据提供方的建议，由于这几个字段的数据是企业选择填写的注册数据，因此这里采取了0填充。



然后将该文件中的9个字段直接按训练集中企业id组合特征，从该文件中总计组合出9个特征，分别为：成立年度、行业大类、注册资本、企业类型，以及5个不明确具体含义的身份指标（MPNUM，INUM，FINZB，FSTINUM，TZINUM）

2.企业变更数据

企业变更数据总计有30W多条数据，但并不是所有公司都有经历变更，因此组合完毕后会出现部分缺失值。因为企业变更次数要大于等于1次，因此该文件中同一企业id可能会对应多条数据，所以处理该文件需要将这多条数据整合后才能加入训练集中。所以这里组合出的第一维特征便是企业的变更次数。然后，该数据总计有4个有效字段，需要我们从这几个字段中组合出有用信息。

（1）针对ALTERNO（变更事项代码）字段，经过统计，该字段总计有12个类别，属于离散型数据，因此直接使用one-hot编码对其编码后组合进训练集中；

（2）针对ALTDATE（变更时间）字段，经过统计，该字段总计有24个类别，因此同上也将其使用one-hot编码后组合进训练集中。除此之外，我们还加入了首次变更日期与最后一次变更日期这两维特征；

（3）关于ALTBE与ALTAF这两个字段数据，由于数据提供方只给了其中一小部分字段的数据，直接将这两个字段求和并组到训练集后分别有90%与98%的缺失率，并且根据实验，去掉这两维特征后，训练集auc值有千分位的提升，因此这里直接抛弃了这两个字段。

综上所述，在企业变更信息数据文件中，我们总计组合出39维特征，分别为：企业变更次数、企业变更事项（one-hot编码后总计12维）、变更时间（one-hot编码后总计24维）、首次变更年度、最后一次变更年度。

3.企业分支信息数据

企业分支数据总计约有11W条数据，因此组合至训练集后必会有缺失数据。同上，一个企业可能有多个分支，所以这里组合出的第一维特征便是企业的分支数目。然后，该数据共有4个有效字段，需要我们从这几个字段中组合出有用信息。

（1）首先是分支id字段，由于这里的id是经过加密的，所以无法才这些id中挖掘信息，因此抛弃这一字段；

（2）针对IFHOME（分支机构是否在同一个省）字段，这里统计了企业分支在省内的概率作为一维特征加入至训练集中；

（3）对于B_REYEAR（分支成立年度）与B_ENDYEAR（分支关停年度）这两个字段，这里统计了分支倒闭概率以及分支存活时间作为两维特征加入至训练集中。

综上所述，在企业分支信息数据中，我们总计组合出4维特征，分别为：企业分支数目，企业分支所处位置与企业同省概率、企业分支倒闭率、企业分支存活时间。

4.企业投资数据

企业投资数据总计约5W多条数据，因此组合至训练集后必会有缺失数据。同上，一个企业可能有多次投资行为，所以这里组合出的第一维特征便是企业投资次数。然后，该数据共有5个有效字段，需要我们从这几个字段中组合出有用信息。

（1）根据IFHOME（被投企业是否在同一个省）字段来得出企业投资外省概率特征；

（2）计算同一投资企业id下BTBL（持股比例）字段总和作为一维特征，并将该特征除以投资次数得到平均每次投股比例特征；

（3）根据BTYEAR（被投企业成立年度）与BTENDYEAR（被投企业停业年度）这两个字段得出企业最近一次投资时间特征以及投资企业倒闭率特征；

（4）由于BTEID（被投企业ID）字段与企业基本信息文件中的EID共享ID，因此我们可以得到训练集中企业被投资次数以及剩余的股份，并将其作为两维特征。

综述所述，在企业投资数据中，我们总计组合出8维特征，分别是：企业投资次数、企业投资外省概率、投资股份总和、平均每次投资股份、企业最近一次投资时间、投资企业倒闭率、企业被投资次数、剩余的股份。

5.企业权利数据

企业权利数据总计有111W条，因此组合至训练集后缺失数据比例较小。同上，一个企业可能有多次权利行为，所以这里组合出的第一维特征便是企业的权利数目。然后，该数据共有4个有效字段，需要我们从这几个字段中组合出有用信息。

（1）针对RIGHTTYPE（权利类型）字段，经过统计，该字段总计有7个类别，属于离散型数据，因此直接使用one-hot编码对其编码后组合进训练集中；

（2）关于TYPECODE(权利ID)字段，该字段仅作为权利唯一标识号，并且是唯一字段，因此没有任何含义，这里直接舍弃；

（3）关于ASKDATE（申请日期）与FBDATE（权利赋予日期）这两个字段，经过分析后，我们提取了最近一次权利赋予日期作为一维特征加入至训练集中。

综上所述，在企业权利数据中，我们总计组合出9维特征，分别是：企业权利数量、企业权利类型（one-hot编码后总计7维）、最近一次权利赋予日期。

6.企业项目数据

企业项目数据总计约3W多条数据，因此组合至训练集后必会有缺失数据。同上，一个企业可能有多个项目，所以这里组合出的第一维特征便是企业项目数量。然后，该数据共有3个有效字段，需要我们从这几个字段中组合出有用信息。

（1）关于TYPECODE（项目ID）字段，该字段仅代表项目唯一标识号，没有实际意义，因此直接舍弃；

（2）关于DJDATE（中标日期）字段，我们首先统计了该字段总共包含24个类别，然而整个项目数据文件只有3W条数据，对比训练集的15W数据，如果直接使用one-hot编码会产生24维特征且这些特征会特别稀疏，对模型训练会产生负面效果，因此这里只取了项目建立时的年份作为划分依据（总计3个年份，分别为2013、2014和2015），最终生成3维特征。然后还取了最新建立的项目的年份作为一维特征；

（3）关于IFHOME（项目地是否是企业登记地）字段，这里同样采取了之前统计项目在外省的概率作为一维特征。

综上所述，在企业项目数据中，我们总计组合出6维特征，分别是：企业项目数量、企业最新项目所在年份、企业项目成立于2013年数目、企业项目成立于2014年数目、企业项目成立于2015年数目、企业项目在外省概率。

7.企业被执行数据

企业项目数据总计约2W多条数据，因此组合至训练集后必会有缺失数据。同上，一个企业可能被处罚多次，所以这里组合出的第一维特征便是企业被处罚次数。然后，该数据共有3个有效字段，需要我们从这几个字段中组合出有用信息。

（1）关于TYPECODE（案件ID）字段，该字段仅代表被执行案件唯一标识号，没有实际意义，因此直接舍弃；

（2）关于LAWDATE(案发日期)字段，经过统计总计有36个类别，然而统计下该文件索引企业id的不重复个数只有5800多条，因此将该数据集整合到训练集后最多只有5800多条数据可能会有该字段数据信息，所以这里就直接取最新被处罚时所在年份作为一维特征；

（3）关于LAWAMOUNT（标的金额（元））字段，这里直接按照企业id取和得到总的罚款金额来作为一维特征加入至训练集中。

综上所述，在企业被执行数据中，我们总计组合出3维特征，分别是：企业被处罚次数、企业最近一次被处罚时间、企业被罚款总金额。

8.企业失信数据

企业项目数据总计约3000多条数据，因此组合至训练集后必会有缺失数据。同上，一个企业可能失信多次，所以这里组合出的第一维特征便是企业失信总次数。然后，该数据共有3个有效字段，需要我们从这几个字段中组合出有用信息。

（1）关于TYPECODE（失信ID）字段，该字段仅代表失信唯一标识号，没有实际意义，因此直接舍弃；

（2）关于FBDATE（失信列入日期）与SXENDDATE（失信结束日期）字段，我们统计了失信结束次数和失信结束概率以及首次失信所在年份作为4维特征。

综上所述，在企业失信数据中，我们总计组合出4维特征，分别是：企业被失信次数、企业首次失信时间、企业最近一次失信时间、失信结束概率。

9.企业招聘数据

企业项目数据总计约3W多条数据，因此组合至训练集后必会有缺失数据。同上，一个企业可能发布多次招聘信息，所以这里组合出的第一维特征便是企业发布招聘信息次数。然后，该数据共有3个有效字段，需要我们从这几个字段中组合出有用信息。

（1）关于WZCODE（招聘网站代码）字段，经过统计，总共有3个类别，属于离散型数据。因此直接采用one-hot编码后组合进训练集中；

（2）关于RECRNUM（招聘职位数量）字段，我们直接按照该文件索引企业id来求和并将该总和除以企业总计发布招聘信息次数从而得到每次招聘所招聘人数的比例，然后将这两维特征组合进训练集中；

（3）关于RECDATE（最近一次招聘日期）字段，这里取了首次招聘时间与最近一次招聘日期这两维特征加入到训练集中。

综上所述，在企业招聘数据中，我们总计组合出8维特征，分别是：企业在1号网站发布招聘次数、企业在2号网站发布招聘次数、企业在3号网站发布招聘次数、企业招聘总人数、企业发布招聘信息次数、平均每次招聘人数、首次招聘时间、最后一次招聘时间。

三、数据清洗

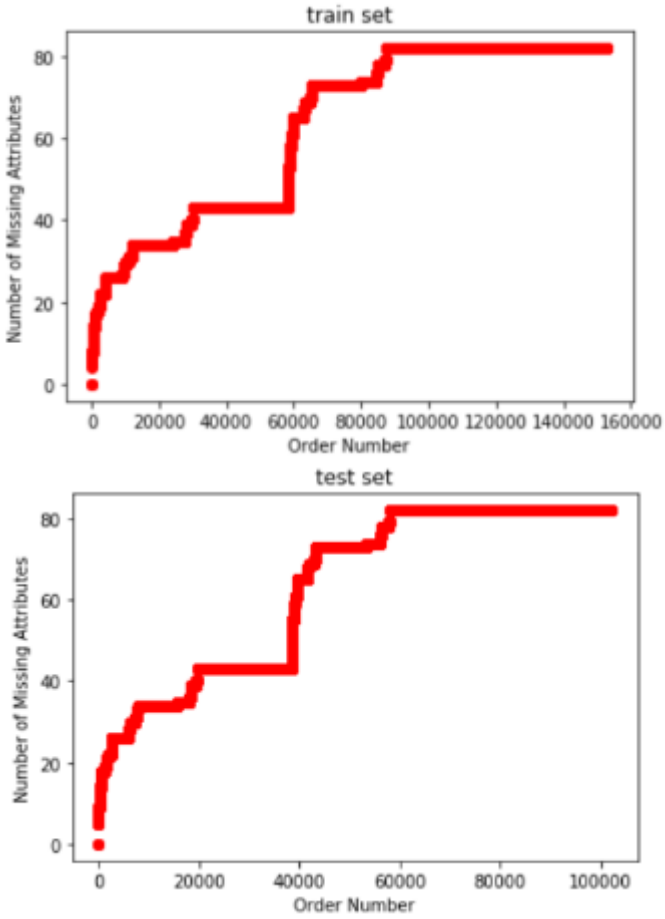
经过上面的步骤，我们才生成我们的训练测试集。而进行模型训练前，还需要对训练集进行检查和清洗，使得其更好的切合算法模型进行训练。

1.按列（属性）统计

在组合训练集的过程中，已经预料到了某些字段会有大量的缺失值，如果按照一般的数据清洗的方法应该是去掉这些特征的，然而经过多次试验，减少上述提到的有意义的特征会对结果产生不良影响，因此在数据清洗时不对特征维度最处理。

2.按行统计

将缺失值个数从小到大排序，以序号为横坐标， 缺失值个数为纵坐标，画出如下散点图



对比 trainset 和testset上的样本的属性缺失值个数，可以发现其分布基本一致，所以没有对其进行处理。

另外，缺失值个数可以作为一个特征，体现企业信息完善度。

四、特征工程

1.类别特征的处理

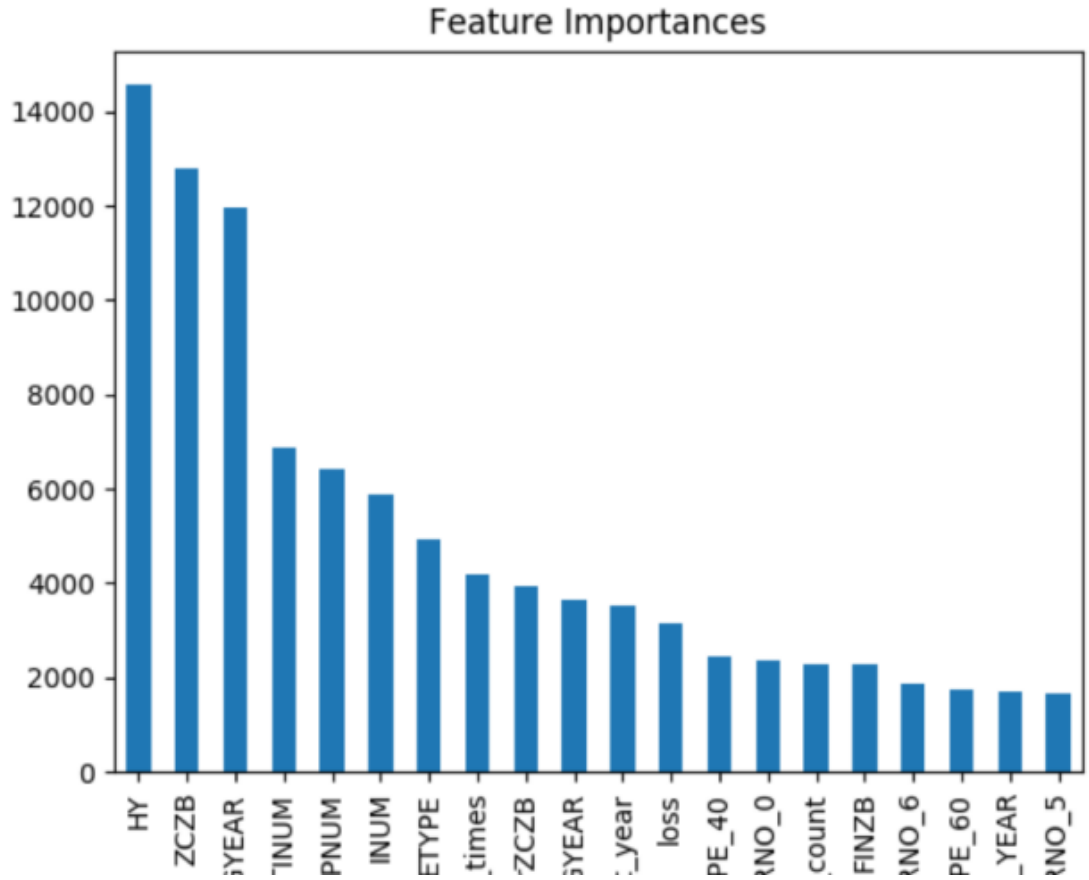
除了组合特征时提到的对某些类别特征进行特殊处理外，其他类别特征都做独热编码。

2.排序特征

对原始特征36维数值型特征按数值从小到大进行排序，得到 36 维排序特征。排序 特征对异常数据有更强的鲁棒性，使得模型更加稳定，降低过拟合的风险。

3.组合特征

Xgboost训练完成后可以输出训练特征的重要性，我们发现这些数据特征的feature score比较大（如下图），即判别性比较高，于是用这部分特征构建了组合特征：将特征两两相除得到 90个特征，然后使用 xgboost 对这 90 多个特征单独训练模型，训练完成后得到特征重要性的排序，取其中 top特征单独训练模型 cv 能达到 0.69+的 AUC 值。将这些特征添加到原始特征体系中，模型 cv 的 AUC 值从 0.6980 提高到 0.7001。另外，也采用了乘法特征（取对数）：log(x*y)，刷选出其中的TOP特征，加入到原始特征体系中，单模型 cv 又提高到了 0.7070 左右。



五、模型设计与分析

经过实验比较，本次最终选择了xgboost作为最终的算法模型。xgboost 是 boosted tree 的一种实现，效率和精度都很高，在各类数据挖掘竞赛中被广泛使用。由于 xgboost 支持分布式运行，很多互联网企业也在实际业务中采用了 xgboost， 实用性很高，比如阿里巴巴的 ODPS 平台部署了 xgboost，腾讯在微信购物里也使用了 xgboost 来做CTR 预估。
基于前面构建出的特征，加上原始特征共有 140 维特征，在这 140 维特征上我们训练 了 xgboost，单模型线下 cross validation 的 auc 值为 0.69800 左右。
随后针对xgboost模型进行了大规模的参数调节，包括调节训练时正负样本的比例，最终调节参数后使得 cross validation 的 auc 值为 0.70780 左右。
由于时间关系，本次项目还未尝试使用模型融合等进一步提高模型准确度的方法，因此本次项目还有一定的提升空间，希望可以再进一步提升预测分类的效果。

六、总结与展望

本文围绕着项目开展一系列的学习，最终的目的是根据现在热门的机器学习与数据挖掘技术搭建出企业经营退出风险预测系统。项目设计如期完成，得到精准度较好结果。由于篇幅的限制，这里就没有详细介绍我们使用的算法——xgboost，如果有需要的话，我们会单独再写一个介绍算法的文档。
主要研究内容有以下几个方面：
(1)对当前互联网中热门技术数据挖掘、机器学习做深入研究，研究发现当前依托热门技术解决社会问题的趋势。同时发现企业经营退出现象，故开始研究预测企业经营退出概率，为风险投资提供建议；
(2)研究并学习机器学习中常见的分类预测算法。主要是逻辑回归、支持向量机、极限梯度提升树等算法；
(3)完成整个项目设计，最终产出预测的文本，同时对比真实的客流量数据，做算法上的偏离率分析；
最后，由于时间的关系，本次项目还未达到极限，还有进一步的提升空间，因此后续还会在多模型融合等方面做进一步研究，争取早日达到最佳的预测效果。