基于机器学习的贷款违约预测

Datawhale September

StarLEE

1. 背景

个人资产业务，即个人信贷业务。主要指运用从负债业务筹集的资金，将资金的使用权在一定期限内有偿让渡给个人，并在贷款到期时收回资金本息以取得收益的业务。个人资产业务作为商业银行主要的收益来源，对商业银行的经营成果起着重要的作用。互联网技术和移动通信技术的不断创新以及与金融行业的快速融合，使得互联网金融中的重要发展模式——网络借贷得到了飞速的发展。网络借贷是指互联网公司利用互联网技术和移动通信技术搭建一个线上平台，为资金提供者和资金需求者提供直接的资金融通服务。这种基于网络平台的借贷方式相较于银行借贷，不仅手续简便、效率高，且允许信用贷款，为自有资本不足、缺乏担保的小额借款人提供了新的出路。它的出现满足了人们多元化的投融资需求，提高了社会闲散资金的利用率。

在以往，商业银行对贷款用户进行信用风险评估时，往往依靠风控人员依赖5C分类法来主观判断，从其个人品格、信用额度、偿付能力、市场经济状况这5个因素对贷款用户进行判断和权衡，以此作为是否贷款给该用户的参考，决定是否发放贷款，这种依赖主观判断的方法显然效率低下，而且评估的时候十分依赖风控人员的主观判断能力，从公司内部管控的角度来看，甚至存在风控人员内部作弊的可能性，不能适应市场经济的快速发展，满足贷款用户的需求，也不能满足网贷平台风险管理的需求。面对数以万计甚至是数以十万计的申请借款的用户时，网贷平台则需要采用各种机器学习的方法来减少监控与检测过程中的人工参与部分，利用自动化的方法提高放款审核的准确率和效率。

目前，网贷行业运用大数据技术进行风险控制管理已经取得了一定的成效，比较成熟的产品有Zest Finance公司所开发的基于数据挖掘和机器学习理论的分析模型，以及美国使用最广泛的个人信用评分系统——FICO信用评分，都是美国借贷行业贷款决策的重要参考标准。而在国内，2014年10月，宜人贷将采用了大数据技术的“极速模式”添加进“宜人贷借款”APP中，积木盒子根据建立的“读秒”标准判断用户的信用等级，拍拍贷于2015年推出的魔镜风控系统被认为是行业内首个基于大数据的风控模型，首个能准确预测借款标的风险概率的风控系统，爱钱进在其两周年发布会上，推出了基于机器学习、深度学习等技术的全新风控体系——“云图动态风控系统”。由此可见，机器学习的蓬勃兴起使网络借贷平台利用多维大数据构建智能风控模型，更加准确的评估个人信用状况，有效地降低违约风险。

2. 数据理解

2.1 变量信息

训练集数据包含了800000个观测值以及47个变量（包括15列匿名变量），变量字段表如下：

**表2.1 变量字段表**

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Description** | **Field** | **Description** |
| id | 为贷款清单分配的唯一信用证标识 | dti | 债务收入比 |
| loanAmnt | 贷款金额 | delinquency\_2years | 借款人过去2年信用档案中逾期30天以上的违约事件数 |
| term | 贷款期限（year） | ficoRangeLow | 借款人在贷款发放时的fico所属的下限范围 |
| interestRate | 贷款利率 | ficoRangeHigh | 借款人在贷款发放时的fico所属的上限范围 |
| installment | 分期付款金额 | openAcc | 借款人信用档案中未结信用额度的数量 |
| grade | 贷款等级 | pubRec | 贬损公共记录的数量 |
| subgrade | 贷款等级之子级 | pubRecBankruptcies | 公开记录清除的数量 |
| employmentTitle | 就业职称 | revolBal | 信贷周转余额合计 |
| employmentLength | 就业年限（年） | revolUtil | 循环额度利用率，或借款人使用的相对于所有可用循环信贷的信贷金额 |
| homeownership | 借款人在登记时提供的房屋所有权状况 | totalAcc | 借款人信用档案中当前的信用额度总数 |
| annuallncome | 年收入 | initialListStatus | 贷款的初始列表状态 |
| verificationStatus | 验证状态 | applicationType | 表明贷款是个人申请还是与两个共同借款人的联合申请 |
| issueDate | 贷款发放月份 | earliesCreditLine | 借款人最早报告的信用额度开立的月份 |
| purpose | 借款人在贷款申请时的贷款用途类别 | title | 借款人提供的贷款名称 |
| postCode | 借款人在贷款申请中提供的邮政编码的前3位数字 | policyCode | 公开可用的策略\_代码=1新产品不公开可用的策略\_代码=2 |
| regionCode | 地区编码 | n系列匿名特征 | 匿名特征n0-n14，为一些贷款人行为计数特征的处理 |

2.2 评价指标

**2.2.1 混淆矩阵 Confuse Matrix**

* True Positive：实例为正，预测为正；
* False Negative：实例为正，预测为负；
* False Positive：实例为负，预测为正；
* True Negative：实例为负，预测为负。

**2.2.2 准确率 Accuracy**

Accuracy不适用于样本分布不均衡的情况（正样本过多或负样本过多）。

**2.2.3 精确率 Precision**

Precision为所有预测为正样本中真实为正的占比。

**2.2.4 召回率 Recall**

Recall为所有正样本中被预测为正的占比。

**2.2.5 F1 Score**

Precision和recall是相互影响的，精确率提高则召回率下降。F1 Score作为精确率和召回率的调和平均数，很好地综合了两者的信息。

**2.2.6 P-R曲线（Precision-Recall Curve）**

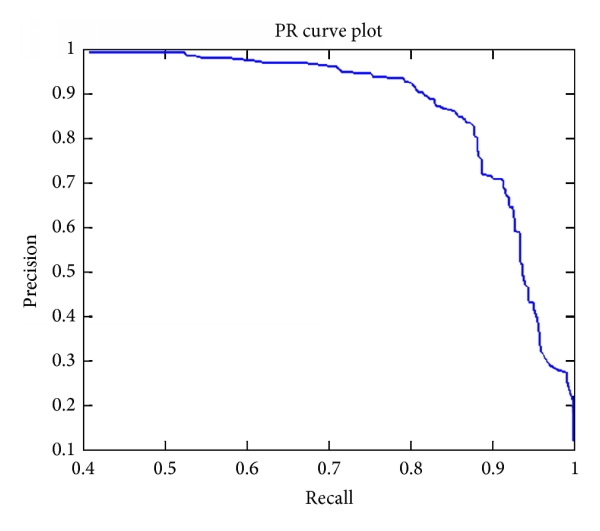


图2.1 P-R曲线示意图

Precision和recall是相互影响的，精确率提高则召回率下降。F1 Score作为精确率和召回率的调和平均数，很好地综合了两者的信息。图2.1展示了P-R曲线样例，可见Precision和Recall存在反相关关系。

**2.2.7 ROC（Receiver Operating Characteristic）**

ROC（receiver operating characteristic）曲线说明了二元分类器系统的鉴别阈值变化时的诊断能力，它是根据一系列不同的二分类方式（分界值或决定阈），以真阳性率（TPR，正样本中预测为正占比）为y轴，假阳性率（FPR，负样本中预测为正占比）为x轴绘制而成，样例如图2.2所示。

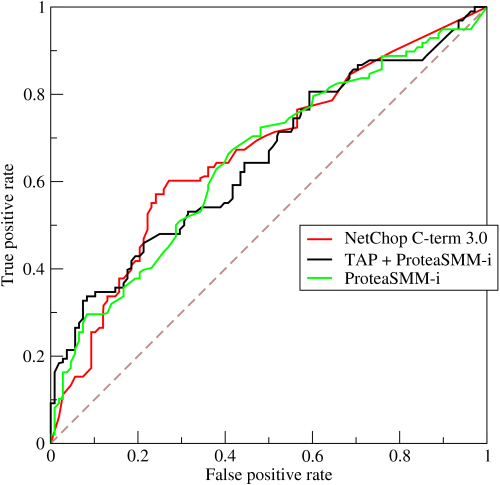


图2.2 ROC示意图

**2.2.8 AUC（Area Under Curve）**

AUC（Area Under Curve）被定义为ROC曲线与坐标轴围成的面积。使用AUC值作为模型的评价标准是因为很多时候ROC曲线并不能清晰地说明哪个分类器效果更好，而作为一个数值，对应AUC更大的分类器效果更好。AUC就是衡量学习器优劣的一种性能指标。

在金融风控领域中，K-S（Kolmogorov-Smirnov）曲线与ROC曲线类似但稍有区别：K-S曲线将TPR和FPR作为y轴，以选定的阈值作为x轴，公式为：；KS值小于0.2认为模型没有区分能力；KS值处于[0.2, 0.3]认为模型有一定区分能力；KS值处于[0.3, 0.5]认为模型有较强区分能力；KS值大于0.75认为模型异常。

2.3 数据读取及评价指标实现

CSV文件数据的读取可由Pandas轻松实现，而评价指标都可以在sklearn中找到已有的api。数据读取结果如图2.3所示，调用api实现评价指标样例结果如图2.4~2.6所示：

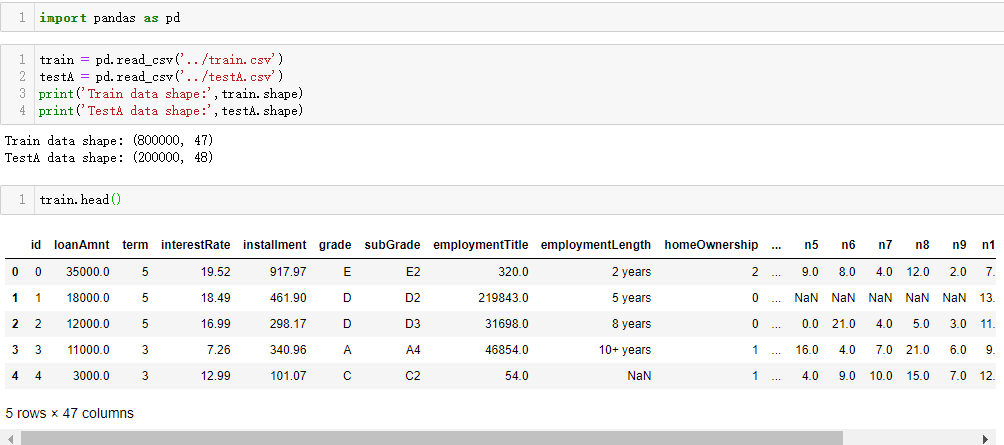


图2.3 训练集及测试集数据读取

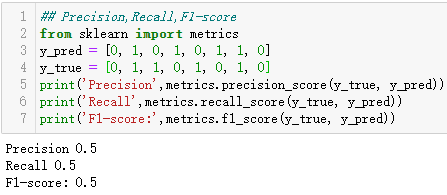
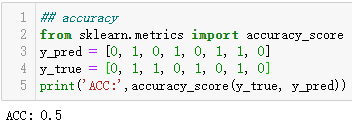
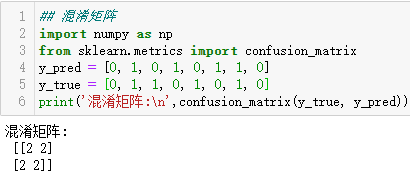


图2.4 混淆矩阵，准确率，精确率，召回率及F1-score调用样例

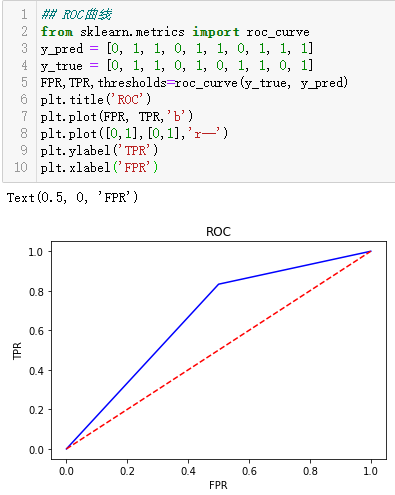
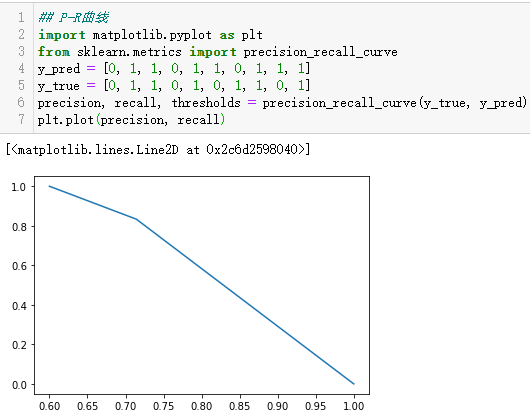


图2.5 P-R，ROC曲线调用样例

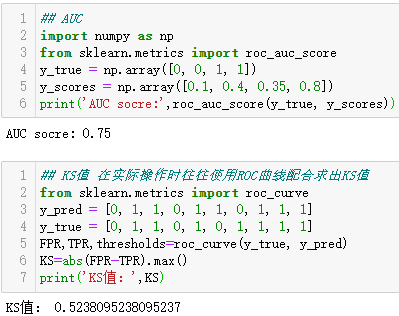


图2.6 AUC，KS值调用样例

3. 数据分析

数据分析是十分重要的，可以根据数据分析得到一定的结论，然后根据这些结论来制定出相应的方案，有的时候根据数据分析可以得到出现错误的原因，可以及时的修正，但是这些前提是，得到的数据必须是真实的，数据分析可以帮助人们做出判断。

3.1 数据总体信息

pandas dataframe 拥有许多attributes和functions，借助 shape, columns, info(), describe() 等可以对数据整体有很好的了解：



图2.7 Training & testing datasets shape

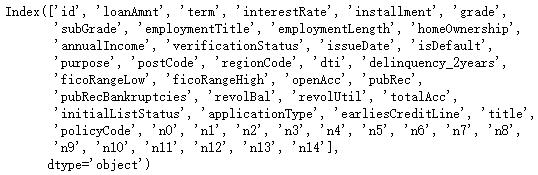


图3.1 Features in dataset

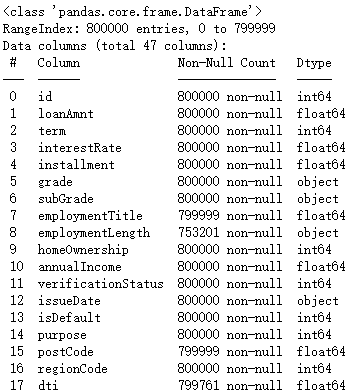


图3.2 Information of dataset

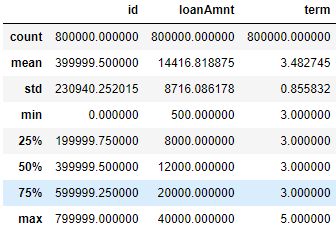


图3.3 Description of features

2.2 缺失值与唯一值

借助 isnull() 函数可得到关于数据是否为空的bool matrix：

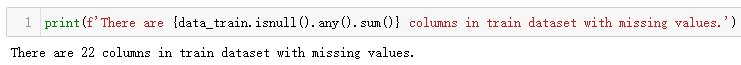


图3.4 包含缺失值features个数统计

数据中尚不存在缺失率大于50%的特征。进一步，还可以对 features 的缺失值情况进行可视化：

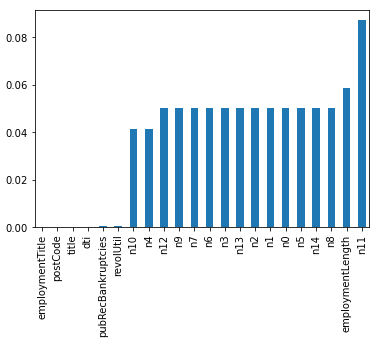


图3.5 Features缺失值统计柱状图

对于缺失值过高的feature，可以认定它对于最终分类的影响是微小的，可以直接删去该feature。其次，通过唯一值分析，发现training和testing数据中存在名为 ‘policyCode’ 的feature具有唯一值，也可直接删去。

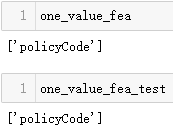


图3.6 Features唯一值分析

通过分析feature的数据类型也可以对数据整体有更好地了解：

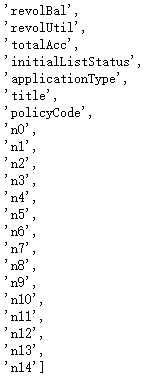
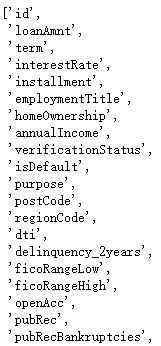


图3.7 数值类型Features



图3.8 类别类型Features

更进一步，对于数值类型feature还可分成连续型与离散型：

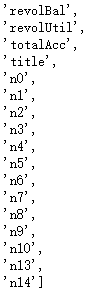
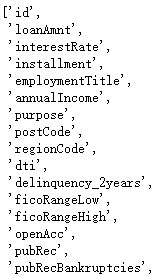


图3.9 连续型Features

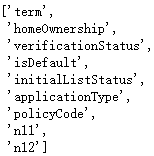


图3.10 离散型Features

借助value\_counts()可以对离散型feature的取值进行分析。对于具有单一值和极差较大的feature可以直接抛去。

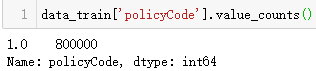


图3.11 具有唯一值的policyCode

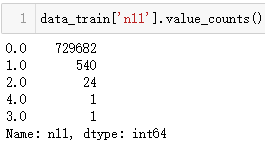


图3.12 具有较大极差的n11

对于连续型数值变量可以对其分布进行可视化

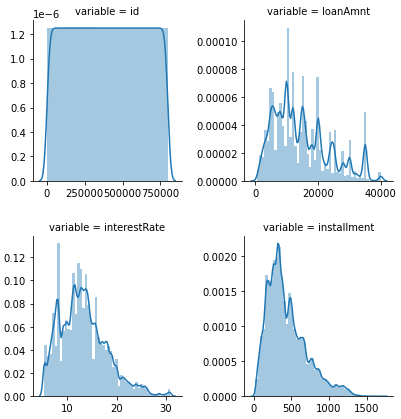


图3.13 连续型features分布

对于不符合正态分布的变量可通过对数log变换，之后再进行观察。一些情形中正态分布可以使得算法收敛速度加快，一些模型更是要求数据具有正太分布，有些只需要数据不要过于具有偏态（可能会影响预测结果）。

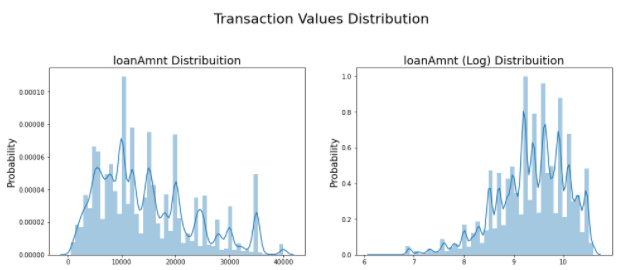


图3.14 对loanAmnt进行log变换

对于非数值型变量，则可直接用value\_counts()进行分析并使用柱状图可视化：

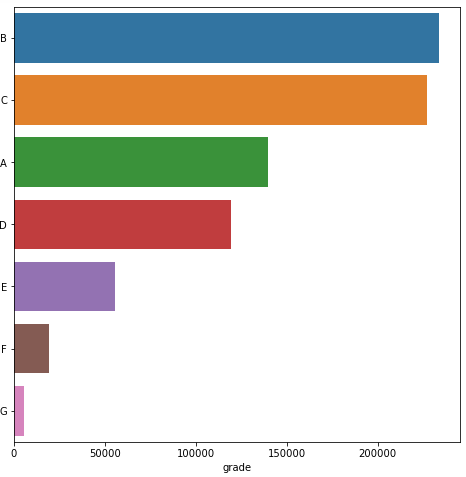
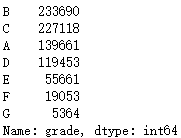


图3.15 对非数值grade变量分析

此外，在training set中还可以根据label的不同对数据进行筛选，分析以及可视化：

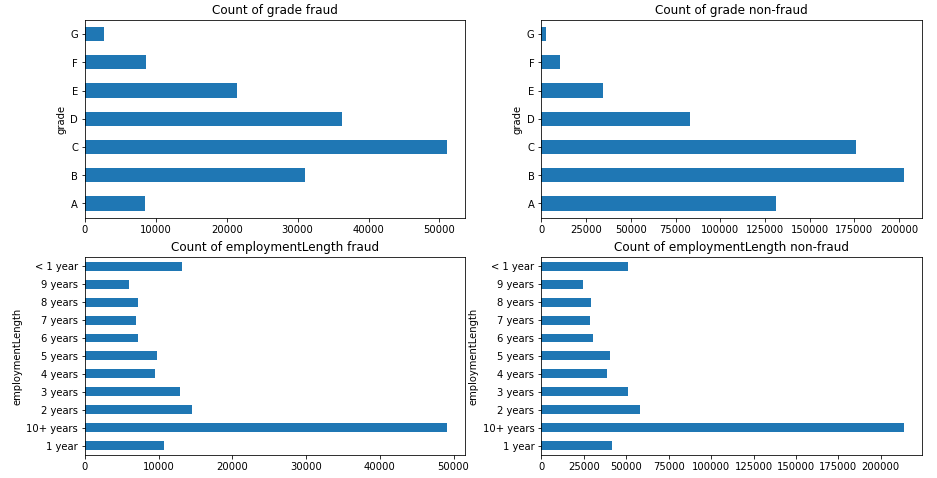


图3.16 基于不同分类的非数值feature数据分析

对于连续变量也可查看对于不同分类，其分布情况

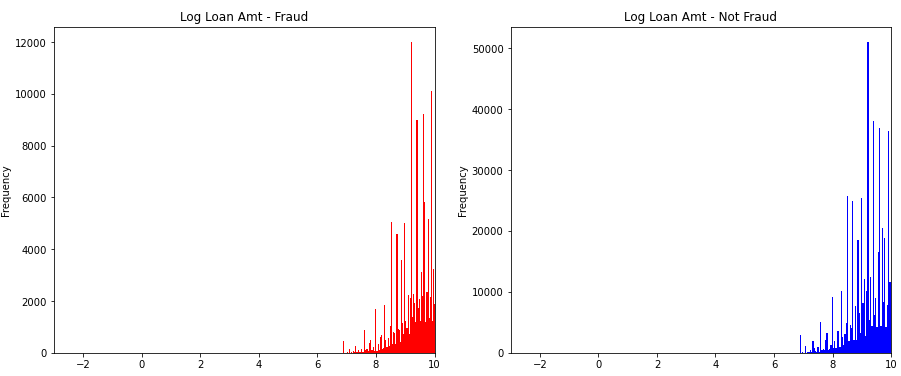


图3.16 基于不同分类的连续型数值feature分布图

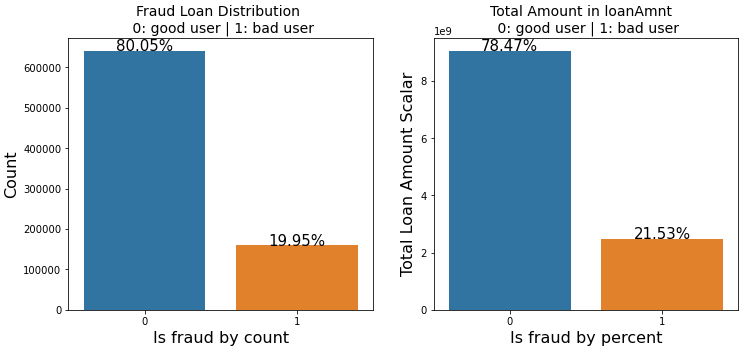


图3.17 基于不同分类的连续型数值feature柱状图

对于时间类型的数据也可以进行可视化和分析：

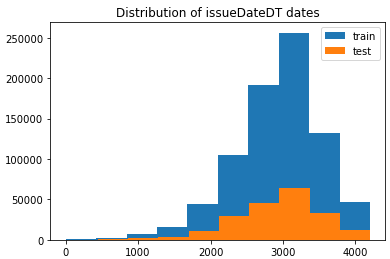


图3.18 时间类型变量issueDateDT可视化

可以看到training和testing data的时间变量是有很大交集的，则该feature很有可能对于最终的分类没有太多贡献。

数据透视表可以动态地改变它们的版面布置，以便按照不同方式分析数据，也可以重新安排行号、列标和页字段。每一次改变版面布置时，数据透视表会立即按照新的布置重新计算数据。另外，如果原始数据发生更改，则可以更新数据透视表。

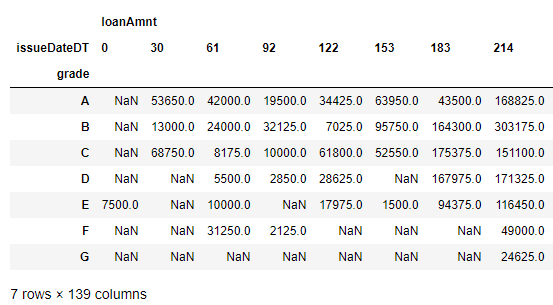


图3.19 数据透视表示例

此外，还可以借助pandas\_profiling来生成dataframe的summary，其中包括了各个变量信息，以及变量间的统计信息。