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：

```
(800000, 47) (200000, 46)
```

图 2.7 Training & testing datasets shape

```
Index(['id', 'loanAmnt', 'term', 'interestRate', 'installment', 'grade',
       'subGrade', 'employmentTitle', 'employmentLength', 'homeOwnership',
       'annualIncome', 'verificationStatus', 'issueDate', 'isDefault',
       'purpose', 'postCode', 'regionCode', 'dti', 'delinquency_2years',
       'ficoRangeLow', 'ficoRangeHigh', 'openAcc', 'pubRec',
       'pubRecBankruptcies', 'revolBal', 'revolUtil', 'totalAcc',
       'initialListStatus', 'applicationType', 'earliesCreditLine', 'title',
       'policyCode', 'n0', 'n1', 'n2', 'n3', 'n4', 'n5', 'n6', 'n7', 'n8',
       'n9', 'n10', 'n11', 'n12', 'n13', 'n14'],
      dtype='object')
```

图 3.1 Features in dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 800000 entries, 0 to 799999
Data columns (total 47 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   id                  800000 non-null  int64
 1   loanAmnt            800000 non-null  float64
 2   term                800000 non-null  int64
 3   interestRate        800000 non-null  float64
 4   installment         800000 non-null  float64
 5   grade               800000 non-null  object
 6   subGrade            800000 non-null  object
 7   employmentTitle     799999 non-null  float64
 8   employmentLength    753201 non-null  object
 9   homeOwnership       800000 non-null  int64
 10  annualIncome        800000 non-null  float64
 11  verificationStatus  800000 non-null  int64
 12  issueDate           800000 non-null  object
 13  isDefault           800000 non-null  int64
 14  purpose             800000 non-null  int64
 15  postCode            799999 non-null  float64
 16  regionCode          800000 non-null  int64
 17  dti                 799761 non-null  float64
```

图 3.2 Information of dataset

	id	loanAmnt	term
count	800000.000000	800000.000000	800000.000000
mean	399999.500000	14416.818875	3.482745
std	230940.252015	8716.086178	0.855832
min	0.000000	500.000000	3.000000
25%	199999.750000	8000.000000	3.000000
50%	399999.500000	12000.000000	3.000000
75%	599999.250000	20000.000000	3.000000
max	799999.000000	40000.000000	5.000000

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：

```
1 print(f'There are {data_train.isnull().any().sum()} columns in train dataset with missing values.')
```

```
There are 22 columns in train dataset with missing values.
```

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：

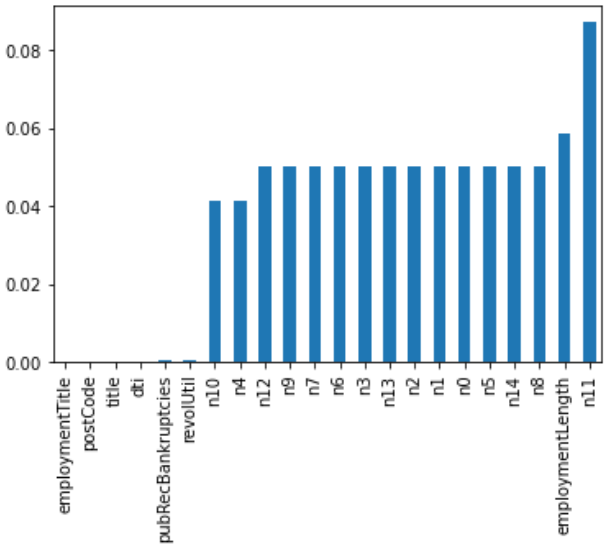

图 3.5 Features 缺失值统计柱状图

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。

```
1 one_value_fea
```

```
['policyCode']
```

```
1 one_value_fea_test
```

```
['policyCode']
```

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：

```
['id',
 'loanAmnt',
 'term',
 'interestRate',
 'installment',
 'employmentTitle',
 'homeOwnership',
 'annualIncome',
 'verificationStatus',
 'isDefault',
 'purpose',
 'postCode',
 'regionCode',
 'dti',
 'delinquency_2years',
 'ficoRangeLow',
 'ficoRangeHigh',
 'openAcc',
 'pubRec',
 'pubRecBankruptcies',
 'revolBal',
 'revolUtil',
 'totalAcc',
 'initialListStatus',
 'applicationType',
 'title',
 'policyCode',
 'n0',
 'n1',
 'n2',
 'n3',
 'n4',
 'n5',
 'n6',
 'n7',
 'n8',
 'n9',
 'n10',
 'n11',
 'n12',
 'n13',
 'n14']
```

图 3.7 数值类型 Features

```
['grade', 'subGrade', 'employmentLength', 'issueDate', 'earliesCreditLine']
```

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：

```
['id',
 'loanAmnt',
 'interestRate',
 'installment',
 'employmentTitle',
 'annualIncome',
 'purpose',
 'postCode',
 'regionCode',
 'dti',
 'delinquency_2years',
 'ficoRangeLow',
 'ficoRangeHigh',
 'openAcc',
 'pubRec',
 'pubRecBankruptcies',
 'revolBal',
 'revolUtil',
 'totalAcc',
 'title',
 'n0',
 'n1',
 'n2',
 'n3',
 'n4',
 'n5',
 'n6',
 'n7',
 'n8',
 'n9',
 'n10',
 'n13',
 'n14']
```

图 3.9 连续型 Features

```
['term',
 'homeOwnership',
 'verificationStatus',
 'isDefault',
 'initialListStatus',
 'applicationType',
 'policyCode',
 'n11',
 'n12']
```

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。

```
1 data_train['policyCode'].value_counts()
```

```
1.0    800000
Name: policyCode, dtype: int64
```

图 3.11 具有唯一值的 policyCode

```
1 data_train['n11'].value_counts()
```

```
0.0    729682
1.0       540
2.0        24
4.0         1
3.0         1
Name: n11, dtype: int64
```

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

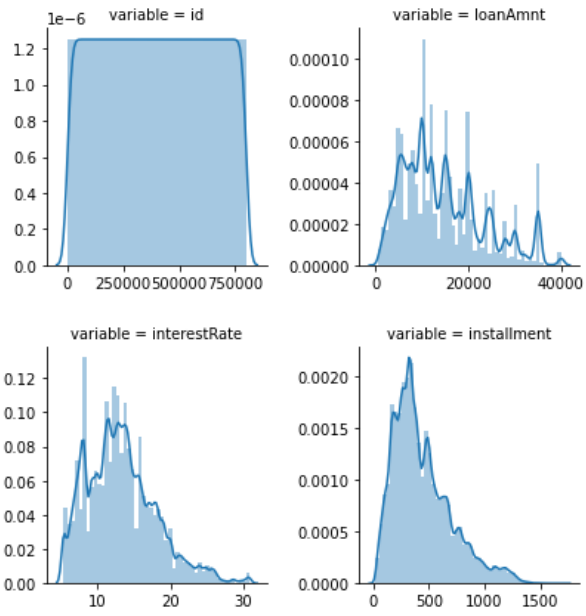

图 3.13 连续型 features 分布

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）。

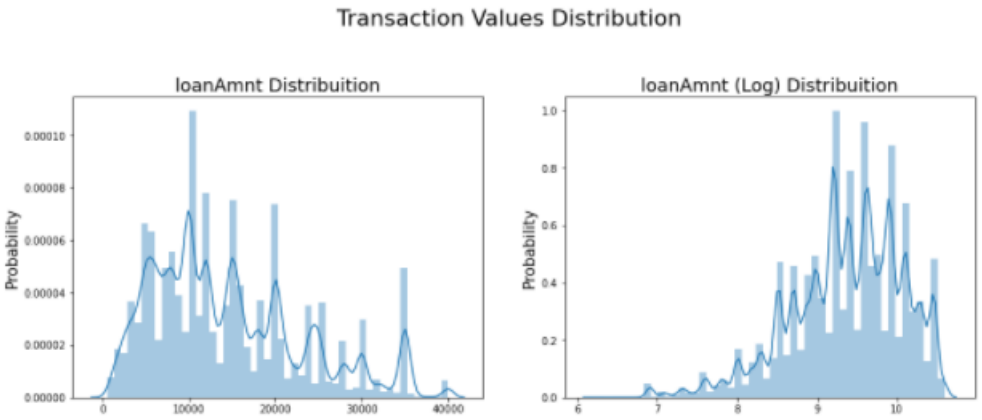

图 3.14 对 loanAmnt 进行 log 变换

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：

```
B    233690
C    227118
A    139661
D    119453
E     55661
F     19053
G      5364
Name: grade, dtype: int64
```

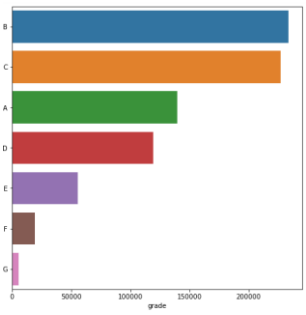

图 3.15 对非数值 grade 变量分析

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：

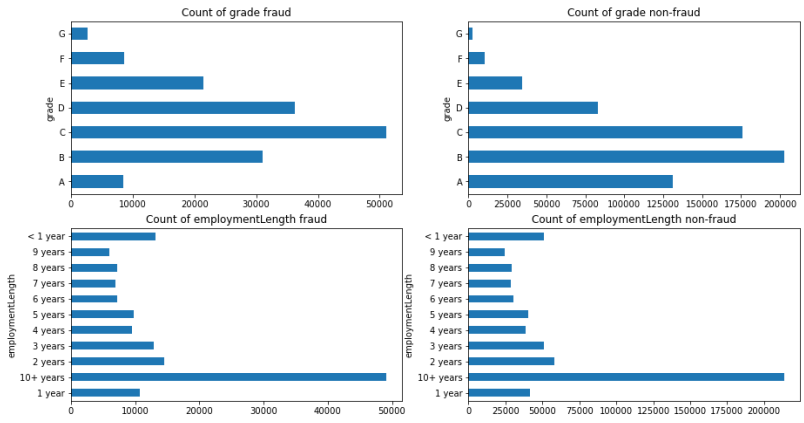

图 3.16 基于不同分类的非数值 feature 数据分析

对于连续变量也可查看对于不同分类，其分布情况

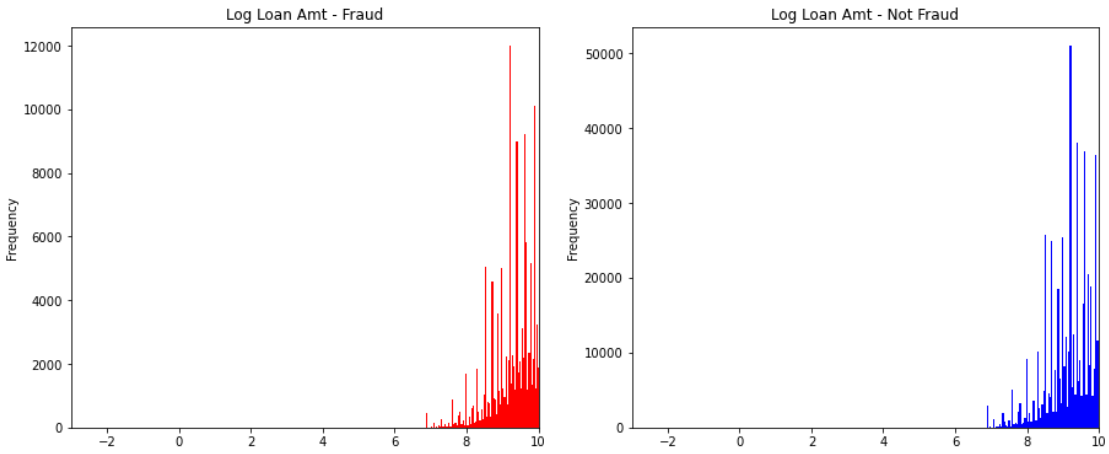

图 3.16 基于不同分类的连续型数值 feature 分布图

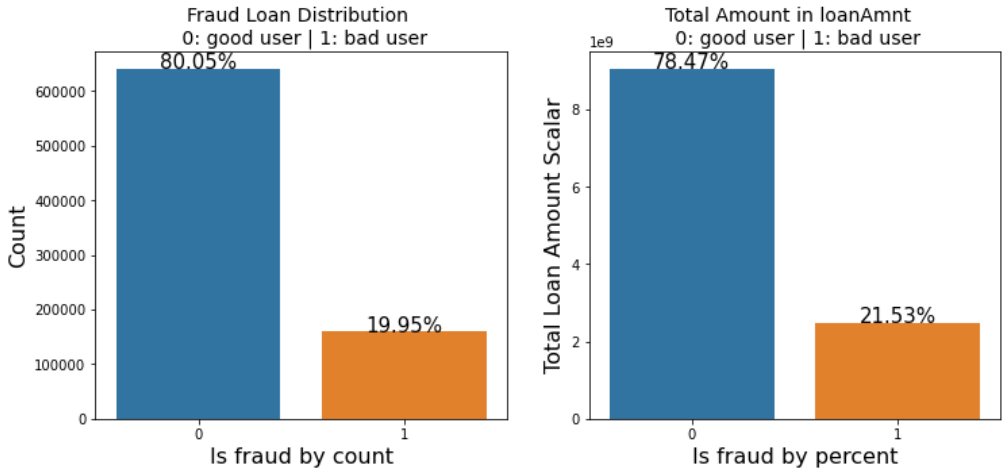

图 3.17 基于不同分类的连续型数值 feature 柱状图

对于时间类型的数据也可以进行可视化和分析：

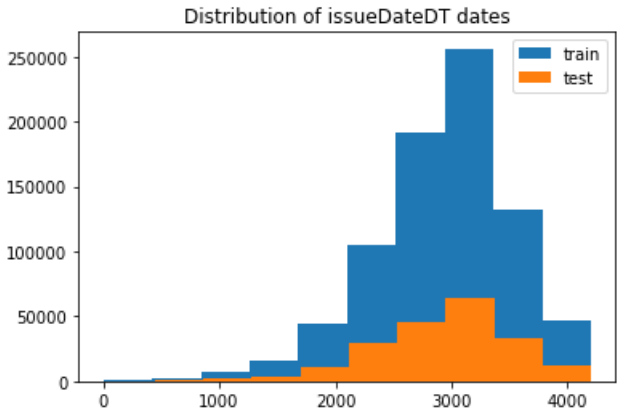

图 3.18 时间类型变量 issueDateDT 可视化

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。

	loanAmnt							
issueDateDT	0	30	61	92	122	153	183	214
grade								
A	NaN	53650.0	42000.0	19500.0	34425.0	63950.0	43500.0	168825.0
B	NaN	13000.0	24000.0	32125.0	7025.0	95750.0	164300.0	303175.0
C	NaN	68750.0	8175.0	10000.0	61800.0	52550.0	175375.0	151100.0
D	NaN	NaN	5500.0	2850.0	28625.0	NaN	167975.0	171325.0
E	7500.0	NaN	10000.0	NaN	17975.0	1500.0	94375.0	116450.0
F	NaN	NaN	31250.0	2125.0	NaN	NaN	NaN	49000.0
G	NaN	NaN	NaN	NaN	NaN	NaN	NaN	24625.0

7 rows × 139 columns

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。