深度之眼
deepshare.net

CCF 个贷违约竞赛 数据EDA和特征工程

导师：Mozak

深度之眼
deepshare.net

目录

深度之眼
deepshare.net

1 数据分析EDA

EDA

Part1 数据分析EDA

EDA

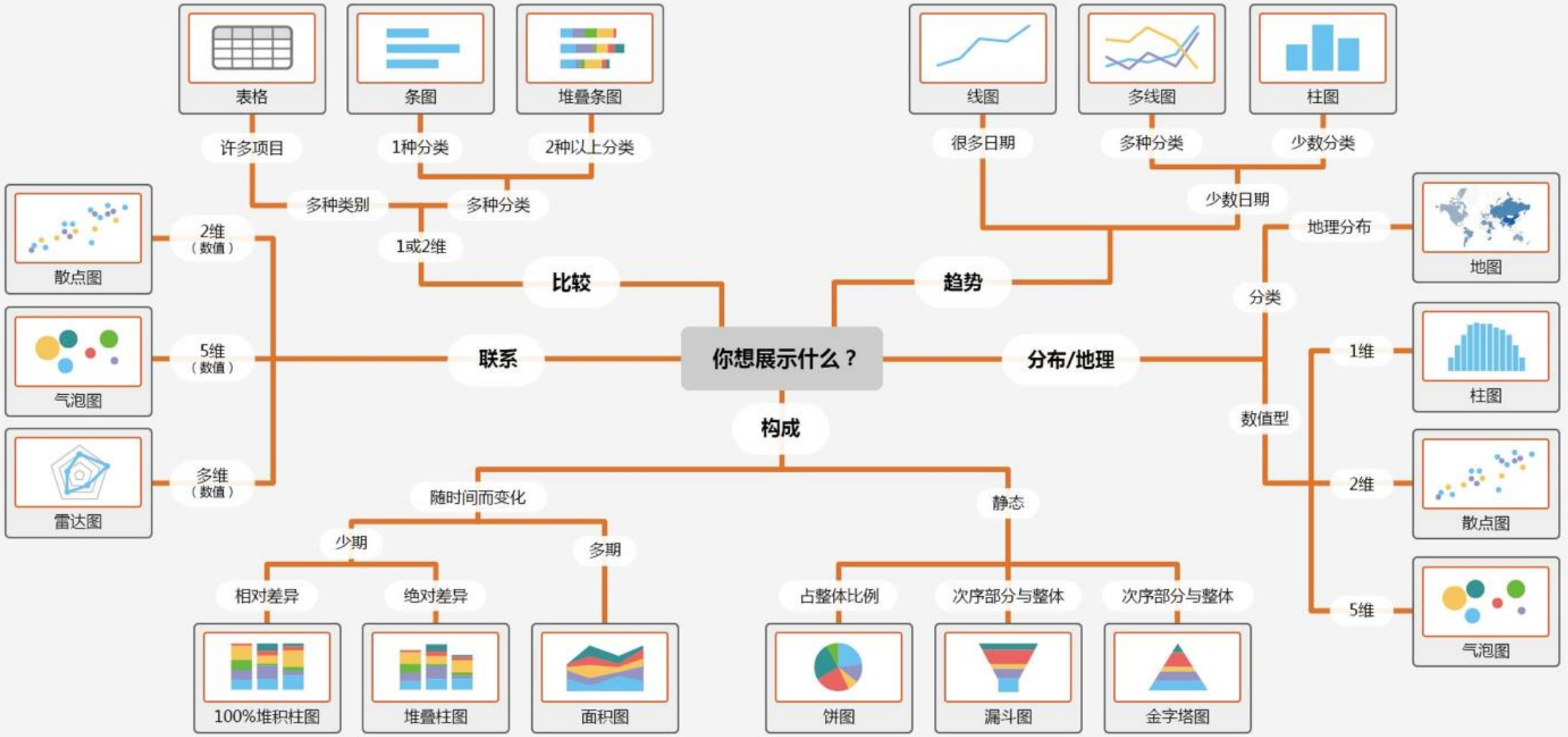
- 数据分析工具：Matplotlib，https://matplotlib.org/

Python环境下最流行的图表库，提供了非常基础且强大的绘图工具，适合定制化开发，是入门Python环境下可视化必备的库。

- 数据分析工具：Seaborn，https://seaborn.pydata.org/

Python环境下常用的可视化图表库，其基于Matplotlib进行二次开发，并对常用的可视化图表进行了定制开发，可快速绘制出漂亮的图表。

如何选择图表的类型？


你想展示什么？
比较
联系
构成
趋势
分布/地理
表格
条图
堆叠条图
线图
多线图
柱图
许多项目
1种分类
2种以上分类
很多日期
多种分类
少数分类
多种类别
多种分类
少数日期
1或2维
2维（数值）
5维（数值）
多维（数值）
散点图
气泡图
雷达图
随时间而变化
静态
少期
多期
相对差异
绝对差异
占整体比例
次序部分与整体
次序部分与整体
100%堆积柱图
堆叠柱图
面积图
饼图
漏斗图
金字塔图
分类
数值型
地理分布
地图
1维
柱图
2维
散点图
5维
气泡图

深度之眼
deepshare.net

2 数据分布分析

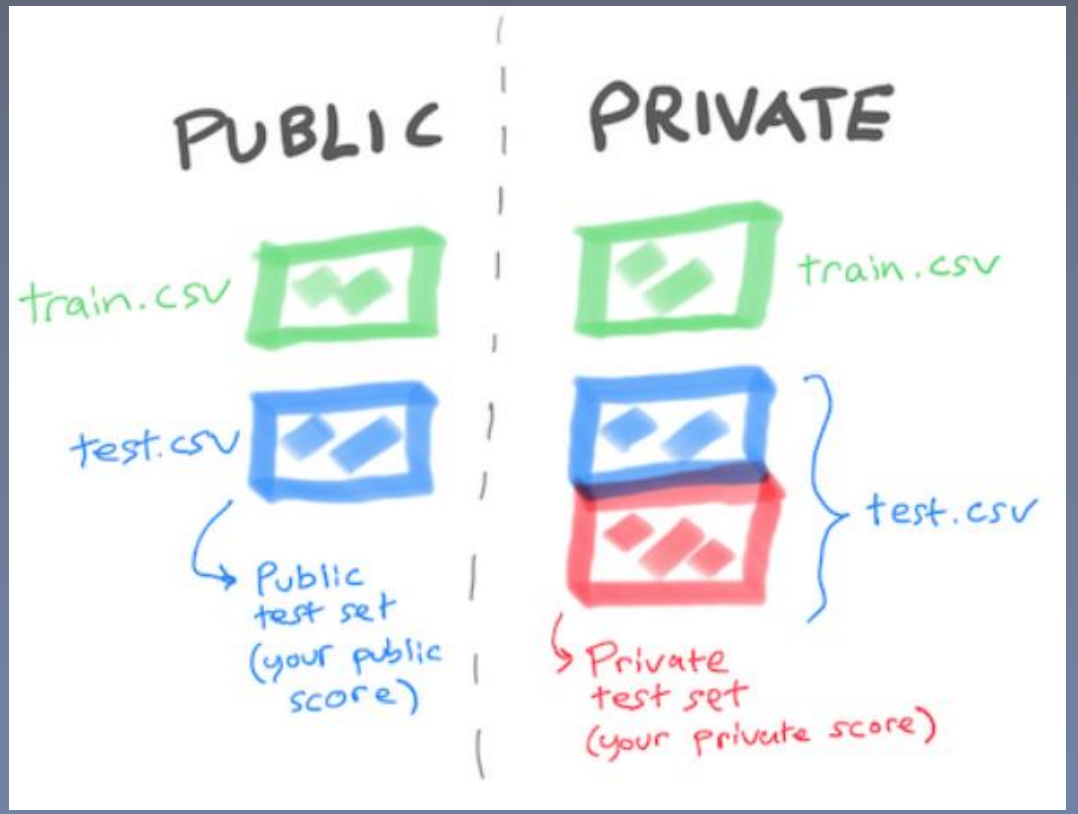
Data Distribution

Part2 数据分布分析

Data Distribution

由于数据集采样和分布的原因导致训练集和线上测试集可能存在分布不一致的情况，也会带来本地交叉验证（Cross Validatation, CV）与线上不一致的情况。

- ✓ 本地得分上分，线上得分下降
- ✓ 本地得分下降，线上得分下降



Part2 数据分布分析

Data Distribution

深度之眼
deepshare.net

Adversarial Validation

构建一个分类模型，分辨训练集和测试集的来源，这里假设使用AUC作为分类精度评价函数。

- ✓ 如果分类模型无法分辨样本（AUC接近0.5），则说明训练集和测试集数据分布比较一致；
- ✓ 如果分类模型可以很好分辨样本（AUC接近1），则说明训练集和测试集数据分布不太一致；

```
train = pd.read_csv( 'data/train.csv' )
test = pd.read_csv( 'data/test.csv' )

train['TARGET'] = 1
test['TARGET'] = 0

data = pd.concat(( train, test ))
x = data.drop( [ 'TARGET', 'ID' ], axis = 1 )
y = data.TARGET



# logistic regression / AUC: 49.82%
# random forest, 10 trees / AUC: 50.05%
# random forest, 100 trees / AUC: 49.95%
```

Part2 数据分布分析

Data Distribution

如果数据分布不一致怎么办？

方法1：假设存在多个训练集，可以使用Adversarial Validation与测试集分布比较一致的一个训练集进行训练；

方法2：假设Adversarial Validation的AUC非常高，可以尝试使用Adversarial Validation选择出与测试集比较相似的样本，构建成为验证集。

方法3：假设数据集可以扩增，则可以使用外部数据来扩增训练数据，以保证训练数据与测试数据的一致性。

深度之眼
deepshare.net

3 特征工程基础

Feature Engineering

Part3 特征工程基础

Feature Engineering

深度之眼
deepshare.net

类别特征（Categorical Features）

✓是最常见的特征：

- ✓个人信息：性别、城市、省份、名族、户口类型等；
- ✓颜色：红色、白色、黑色、粉色等；
- ✓国家：中国、美国、英国、新加坡等；
- ✓动物：猫、狗、蛇、老虎、猴等；

✓任何时候都需要进行处理的数据；

✓高基数（High cardinality）会带来离散数据；

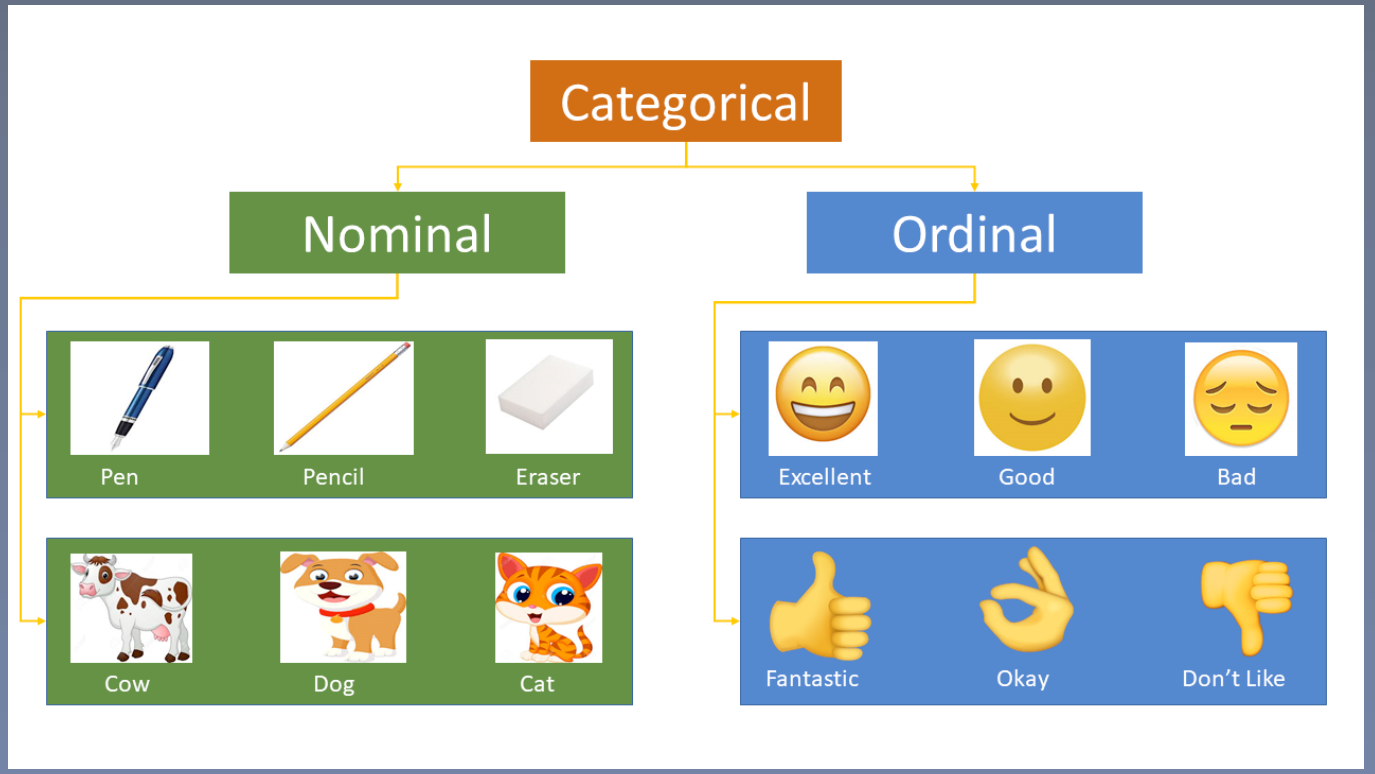
✓很难进行缺失值填充；

Part3 特征工程基础

Feature Engineering

类别特征（Categorical Features）

可分类两种类型：无序（Nominal） 和 有序（Ordinal）



Part3 特征工程基础

Feature Engineering

深度之眼
deepshare.net

```
df = pd.DataFrame({
    'student_id': [1,2,3,4,5,6,7],
    'country': ['China', 'USA', 'UK', 'Japan', 'Korea', 'China', 'USA'],
    'education': ['Master', 'Bachelor', 'Bachelor', 'Master', 'PHD', 'PHD', 'Bachelor'],
    'target': [1, 0, 1, 0, 1, 0, 1]
})
df.head(10)
```

executed in 14ms, finished 13:19:51 2020-08-02

	student_id	country	education	target
0	1	China	Master	1
1	2	USA	Bachelor	0
2	3	UK	Bachelor	1
3	4	Japan	Master	0
4	5	Korea	PHD	1
5	6	China	PHD	0
6	7	USA	Bachelor	1

Part3 特征工程基础

Feature Engineering

深度之眼
deepshare.net

类别特征（Categorical Features）编码方式：

✓ One Hot Encoding（独热编码）

- 形式：编码为One-of-K的K维向量形式；
- 用途：在所有的线性模型；
- 优点：简单，能够将类别特征进行有效编码；
- 缺点：会带来维度爆炸和特征稀疏；
- 实现方法：
  - 在pandas中使用get_dummies；
  - 在sklearn中使用OneHotEncoder；

```
pd.get_dummies(df, columns=['education'])
```

executed in 14ms, finished 13:21:27 2020-08-02

	student_id	country	target	education_Bachelor	education_Master	education_PHD
0	1	China	1	0	1	0
1	2	USA	0	1	0	0
2	3	UK	1	1	0	0
3	4	Japan	0	0	1	0
4	5	Korea	1	0	0	1
5	6	China	0	0	0	1
6	7	USA	1	1	0	0

Part3 特征工程基础

Feature Engineering

深度之眼 deepshare.net

类别特征（Categorical Features）编码方式：

✓Label Encoding（标签编码）

- 形式：将每个类别变量使用独立的数字ID编码
- 用途：在树模型中比较适合；
- 优点：简单，不增加类别的维度；
- 缺点：会改变原始标签的次序关系；
- 实现方法：
  - pandas中的facotrize
  - sklearn中的LabelEncoder

```
df['country_LabelEncoder'] = pd.factorize(df['country'])[0]
df.head(10)
```

executed in 10ms, finished 13:34:47 2020-08-02

	student_id	country	education	target	country_LabelEncoder
0	1	China	Master	1	0
1	2	USA	Bachelor	0	1
2	3	UK	Bachelor	1	2
3	4	Japan	Master	0	3
4	5	Korea	PHD	1	4
5	6	China	PHD	0	0
6	7	USA	Bachelor	1	1

```
pd.factorize(df['country'])
```

executed in 7ms, finished 13:34:49 2020-08-02

```
(array([0, 1, 2, 3, 4, 0, 1]),
 Index(['China', 'USA', 'UK', 'Japan', 'Korea'], dtype='object'))
```

Part3 特征工程基础

Feature Engineering

深度之眼
deepshare.net

类别特征（Categorical Features）编码方式：

✓Ordinal Encoding（顺序编码）

- 形式：按照类别大小关系进行编码
- 用途：在大部分场景都适用；
- 优点：简单，不增加类别的维度；
- 缺点：需要人工知识，且对未出现的数值不友好；
- 实现方法：手动定义字典映射；

```
df['education'] = df['education'].map(
                    {'Bachelor': 1,
                    'Master': 2,
                    'PHD': 3})
df.head(10)
```

executed in 11ms, finished 13:46:25 2020-08-02

	student_id	country	education	target
0	1	China	2	1
1	2	USA	1	0
2	3	UK	1	1
3	4	Japan	2	0
4	5	Korea	3	1
5	6	China	3	0
6	7	USA	1	1

Part3 特征工程基础

Feature Engineering

深度之眼 deepshare.net

类别特征（Categorical Features）编码方式：

✓Frequency Encoding、Count Encoding

- 形式：将类别出现的次数或频率进行编码
- 用途：在大部分情况下都通用
- 优点：简单，可以统计类别次数；
- 缺点：容易受到类别分布带来的影响；
- 实现方法：使用次数统计；

```
df['country_count'] = df['country'].map(df['country'].value_counts()) / len(df)
df.head(10)
```

executed in 22ms, finished 14:36:42 2020-08-02

	student_id	country	education	target	country_count
0	1	China	Master	1	0.285714
1	2	USA	Bachelor	0	0.285714
2	3	UK	Bachelor	1	0.142857
3	4	Japan	Master	0	0.142857
4	5	Korea	PHD	1	0.142857
5	6	China	PHD	0	0.285714
6	7	USA	Bachelor	1	0.285714

```
df['country_count'] = df['country'].map(df['country'].value_counts())
df.head(10)
```

executed in 13ms, finished 14:36:43 2020-08-02

	student_id	country	education	target	country_count
0	1	China	Master	1	2
1	2	USA	Bachelor	0	2
2	3	UK	Bachelor	1	1
3	4	Japan	Master	0	1
4	5	Korea	PHD	1	1
5	6	China	PHD	0	2
6	7	USA	Bachelor	1	2

Part3 特征工程基础

Feature Engineering

深度之眼
deepshare.net

类别特征（Categorical Features）编码方式：

✓Mean/Target Encoding

- 形式：将类别对应的标签概率进行编码；
- 用途：在大部分场景都可以通用；
- 优点：让模型更容易学习标签信息；
- 缺点：容易过拟合；
- 实现方法：使用次数统计；

```
df['country_target'] = df['country'].map(df.groupby(['country'])['target'].mean())
df.head(10)
```

executed in 12ms, finished 14:49:42 2020-08-02

	student_id	country	education	target	country_target
0	1	China	Master	1	0.5
1	2	USA	Bachelor	0	0.5
2	3	UK	Bachelor	1	1.0
3	4	Japan	Master	0	0.0
4	5	Korea	PHD	1	1.0
5	6	China	PHD	0	0.5
6	7	USA	Bachelor	1	0.5

Part3 特征工程基础

Feature Engineering

深度之眼
deepshare.net

数值特征（Numerical Features）

✓是常见的连续特征：

- ✓年龄：18、19、25、40；
- ✓成绩：55、60、75、80、95；
- ✓经纬度：45.87、23.89、21.21；

✓容易出现异常值和离群点；

Part3 特征工程基础

Feature Engineering

数值特征（Numerical Features）编码方式：

✓Round

- 形式：将数值进行缩放、取整；
- 用途：在大部分场景都可以通用；
- 优点：可以保留数值大部分信息；
- 缺点：
- 实现方法：

```
df['age_round1'] = df['age'].round()
df['age_round2'] = (df['age'] / 10).astype(int)
df.head(10)
```

executed in 14ms, finished 15:20:54 2020-08-02

	student_id	country	education	age	target	age_round1	age_round2
0	1	China	Master	34.5	1	34.0	3
1	2	USA	Bachelor	28.9	0	29.0	2
2	3	UK	Bachelor	19.5	1	20.0	1
3	4	Japan	Master	23.6	0	24.0	2
4	5	Korea	PHD	19.8	1	20.0	1
5	6	China	PHD	29.8	0	30.0	2
6	7	USA	Bachelor	31.7	1	32.0	3

Part3 特征工程基础

Feature Engineering

数值特征（Numerical Features）编码方式：

✓Binning

- 形式：将数值进行分箱；
- 用途：在大部分场景都可以通用；
- 优点：可以将连续特征离散化
- 缺点：
- 实现方法：

```
df['age_<20'] = (df['age'] <= 20).astype(int)
df['age_20-25'] = ((df['age'] > 20) & (df['age'] <=25)).astype(int)
df['age_20-25'] = ((df['age'] > 25) & (df['age'] <= 30)).astype(int)
df['age_>30'] = (df['age'] > 30).astype(int)
df.head(10)
```

executed in 16ms, finished 15:26:29 2020-08-02

	student_id	country	education	age	target	age_<20	age_20-25	age_>30
0	1	China	Master	34.5	1	0	0	1
1	2	USA	Bachelor	28.9	0	0	1	0
2	3	UK	Bachelor	19.5	1	1	0	0
3	4	Japan	Master	23.6	0	0	0	0
4	5	Korea	PHD	19.8	1	1	0	0
5	6	China	PHD	29.8	0	0	1	0
6	7	USA	Bachelor	31.7	1	0	0	1

深度之眼
deepshare.net

4 比赛上分思路

The Right Way

4 比赛进化路线

The Right way

Kaggle比赛**最关键最正确**的路线是：找到一份代码，线上得分与线下验证趋势相同；

- ✓ 线下如何验证，线上如何评分?
- ✓ 线下得分上分，线上如何变化?

你需要做的事情：

- ✓ 构建比赛线下验证集；
- ✓ 对特征进行编码、并构造新特征；
- ✓ 完成模型集成；
- ✓ 多个数据集完成迁移学习；

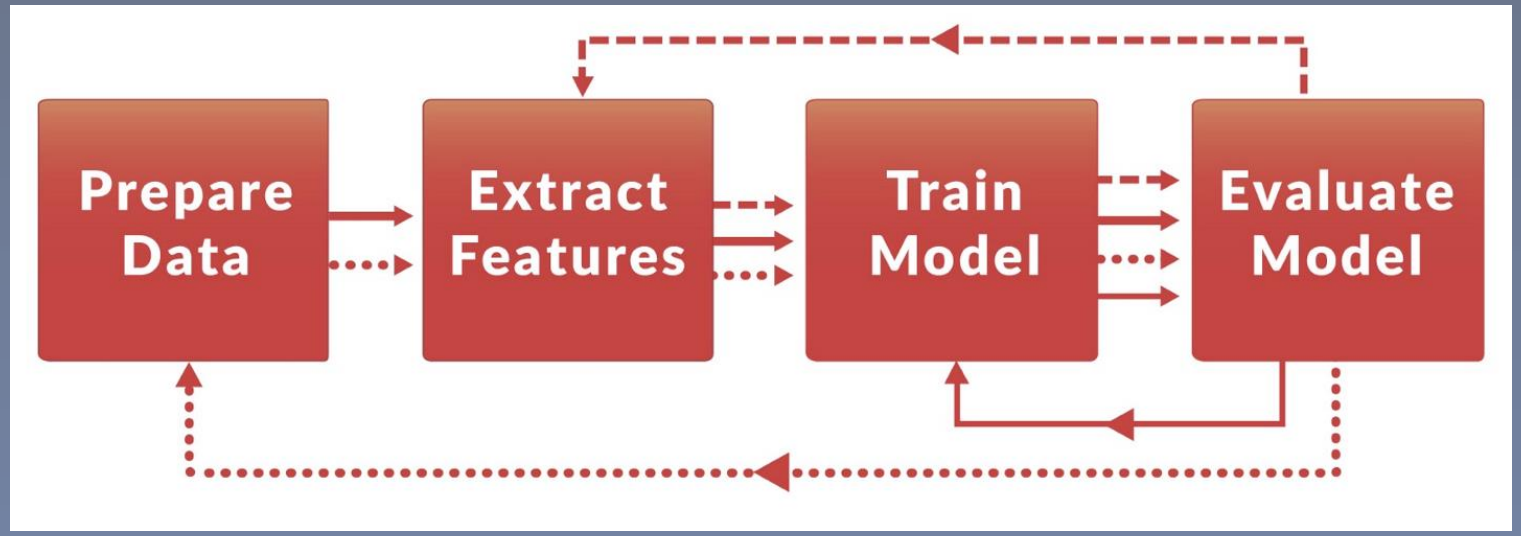
4 比赛进化路线

The Right way

深度之眼
deepshare.net

完整的竞赛流程是什么?

- ✓ 结构化：70%特征工程 + 20%模型 + 10%集成学习；
- ✓ 非结构化：40%模型选择 + 30%数据扩增 + 10模型调参%



请让我们一起立一个flag！

我承诺：

4周努力上TOP100！

结语

再小的细节，也值得被认真对待