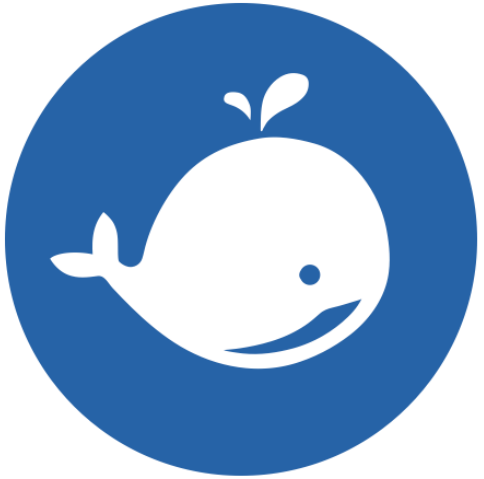


Datawhale

开源学习社区

Part1 数据竞赛是什么？

What is Data Competition ?

数据竞赛是以工业/学术**问题为导向**的，聚合广泛、跨学科人才参与的，利用数据研发算法模型和探索解决方案的研发模式。

数据竞赛是：

- ✓一种众包的竞赛模型（对参赛人员门槛没有限制）；
- ✓有明确的问题背景（具有较强的业务或数据背景）；

	数据竞赛	普通学科竞赛
内容	数据挖掘 & 机器学习	学科知识点
打分反馈	支持	不支持
定量打分	支持	不支持
交互式	支持	不支持

Part1 数据竞赛是什么？

What is Data Competition ?

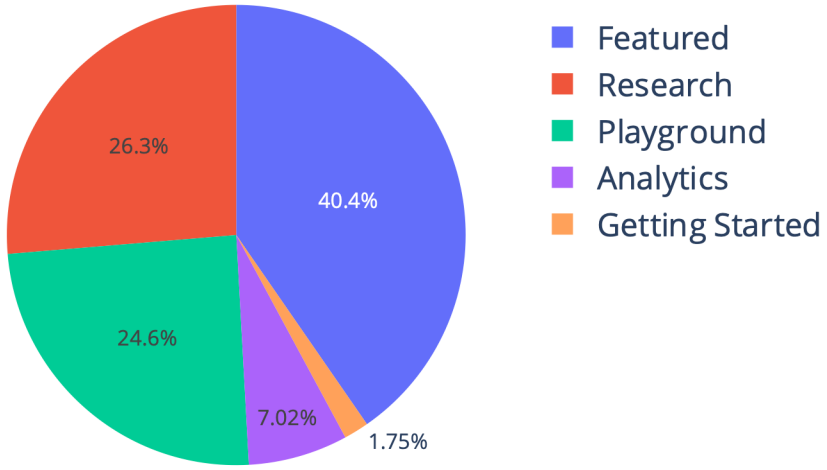
■ 数据竞赛平台：Kaggle/天池/DC竞赛/DataFountain/FlyAI等

Kaggle是全球最大的数据竞赛平台，每年会举办几十场竞赛，主要以算法赛和可视化比赛居多。Kaggle具有完整的比赛平台机制：从赛题介绍、数据分析、评分、排名和最终的分享。

Part1 数据竞赛是什么？

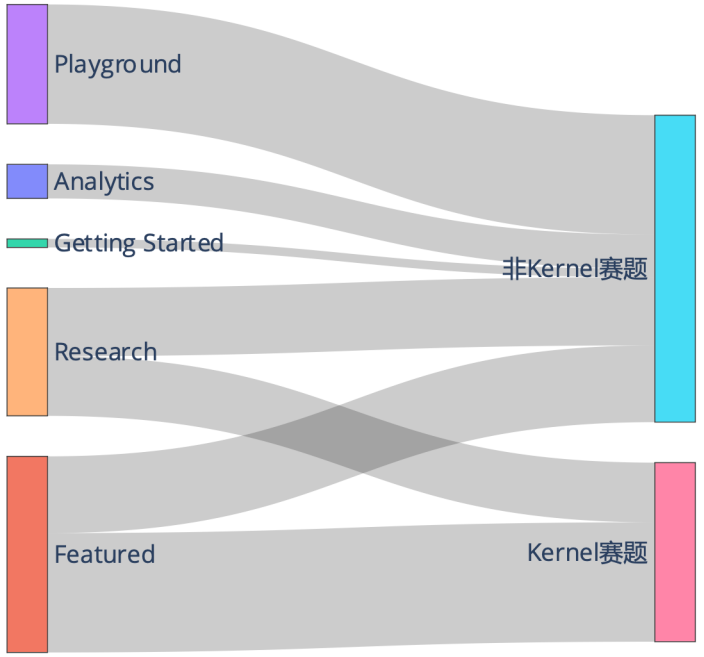
What is Data Competition ?

- Kaggle上数据竞赛有哪些类型？



- Feature： 工业赛赛题，难度较大
- Research：学术赛题，难度较大
- Playground：练习赛，难度适中
- Analytics：数据分析赛
- Getting Started：入门赛，难度较低

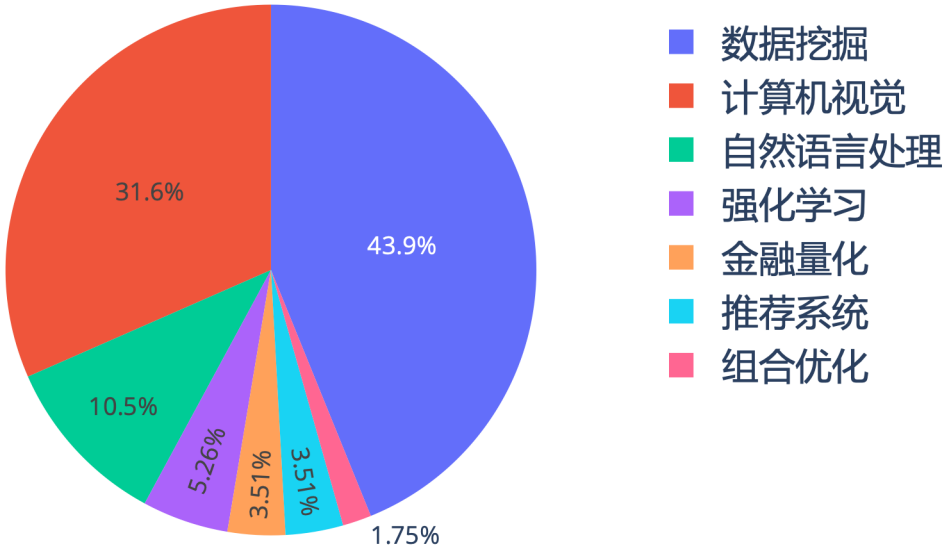
- Kernel赛题：需通过Notebook提交的比赛
- 非Kernel赛题：通过Notebook & 文件提交的比赛

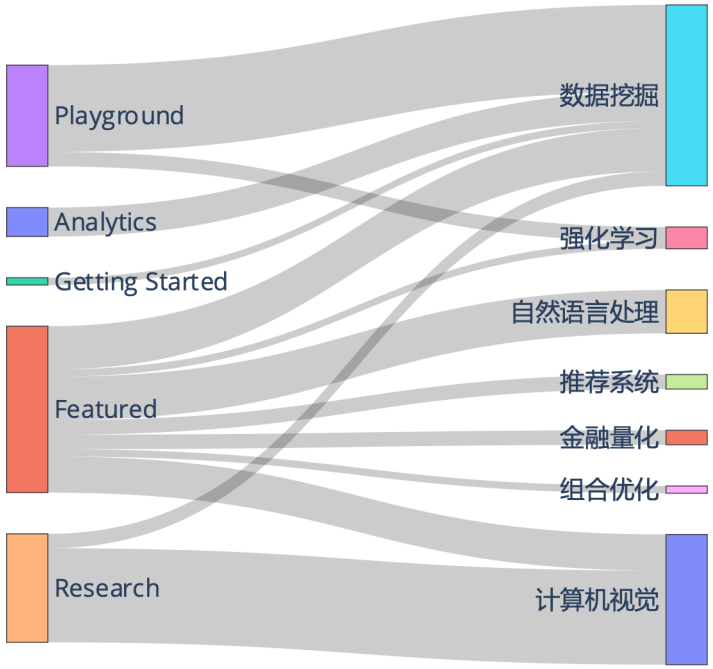


Part1 数据竞赛是什么？

What is Data Competition ?

- Kaggle上数据竞赛有哪些类型？





Part2 竞赛基础知识

Basic knowledge

为什么要做数据分析？

- ✓ 分析数据的质量、噪音；
- ✓ 分析字段的类型、取值、分布，为后续操作提供依据；
- ✓ 分析字段的含义、相关性；

如何做数据分析？

- ✓ 统计；
- ✓ 可视化；

Part2 竞赛基础知识

Basic knowledge

建模前做数据分析：

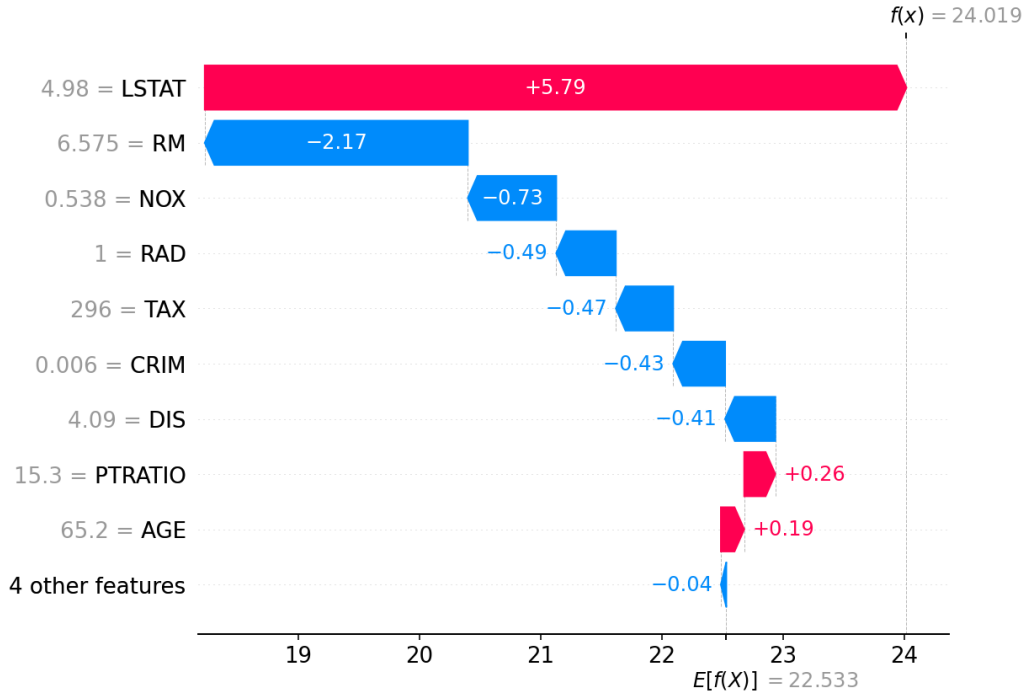
✓ 为建模细节提供参考；

✓ 为模型选择提供参考；

建模后做数据分析：

✓ 特征重要性分析；

✓ 误差分析；



https://github.com/slundberg/shap

Part2 竞赛基础知识

Basic knowledge

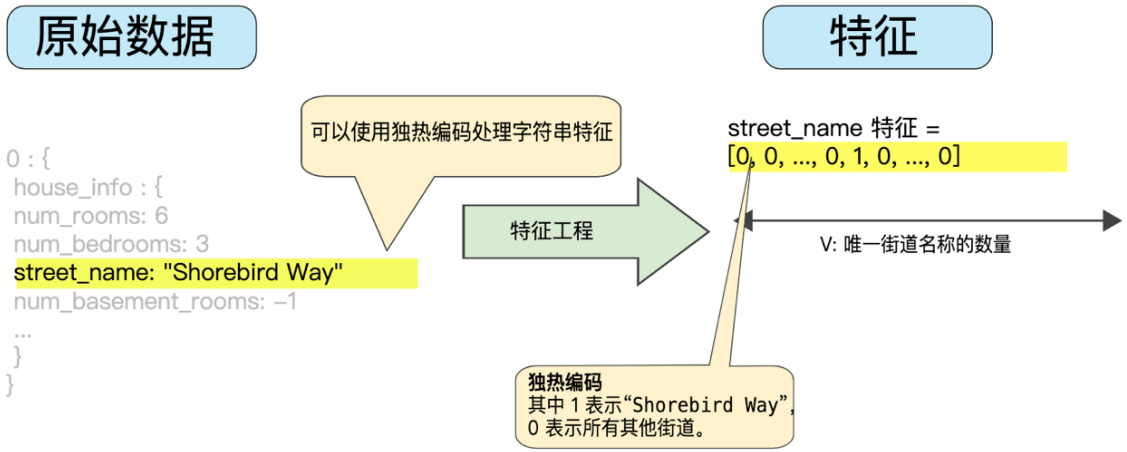
■ 特征工程是什么？

特征工程是原始数据转变为模型的训练数据的过程，就是获取更好的训练数据特征，使得机器学习模型逼近这个上限。特征工程是数据挖掘中重要的部分，包括特征构建、特征提取、特征选择三个部分。

□ 为什么需要特征工程？

✓ 数据的原始字段并不直接适合送给模型训练

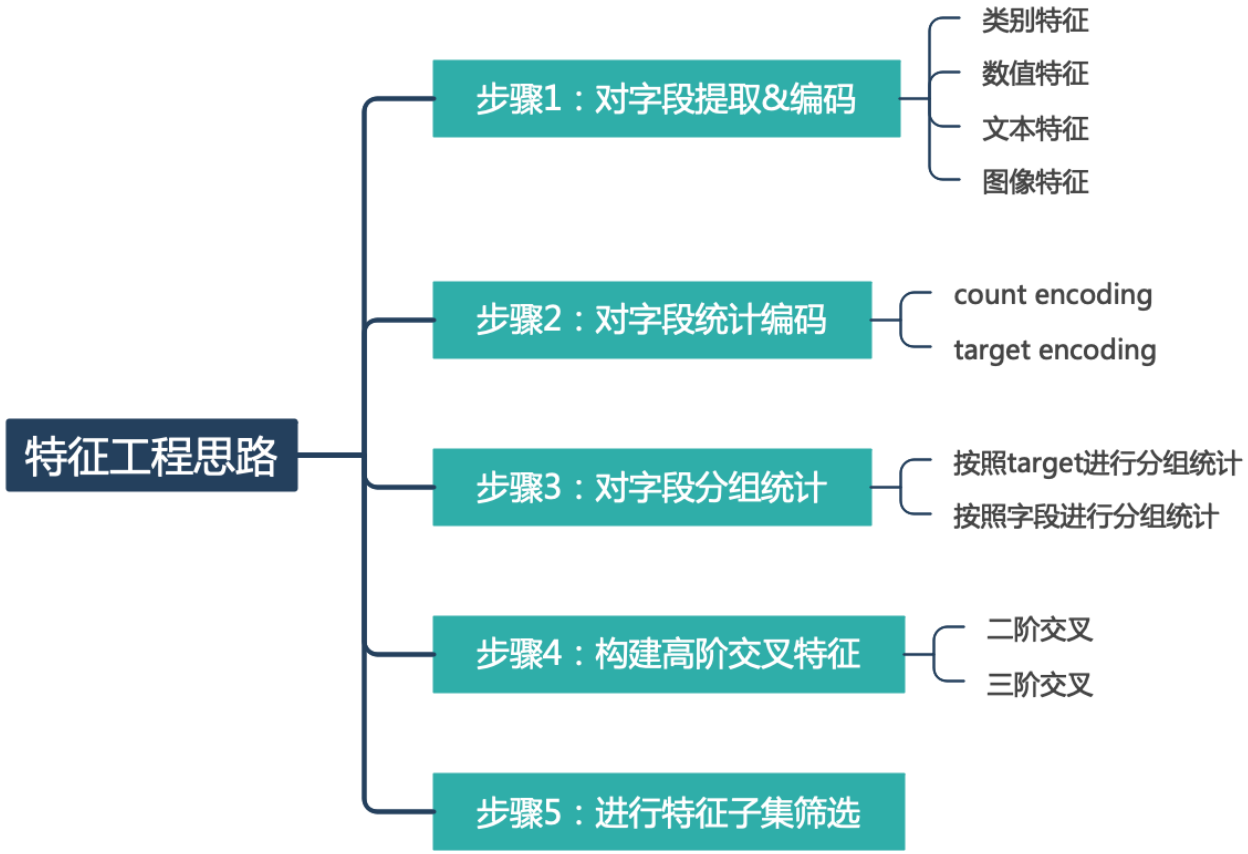
✓ 数据的原始字段并不能体现数据内在的含义



Part2 竞赛基础知识

Basic knowledge

- 如何做特征工程？
  - ✓ 字段如何编码？
  - ✓ 如何构造新特征？
  - ✓ 如何筛选特征？



Part3 入门结构化竞赛

Tableaur dataset

结构化数据是以行列存储的数据，以表格形式存储。

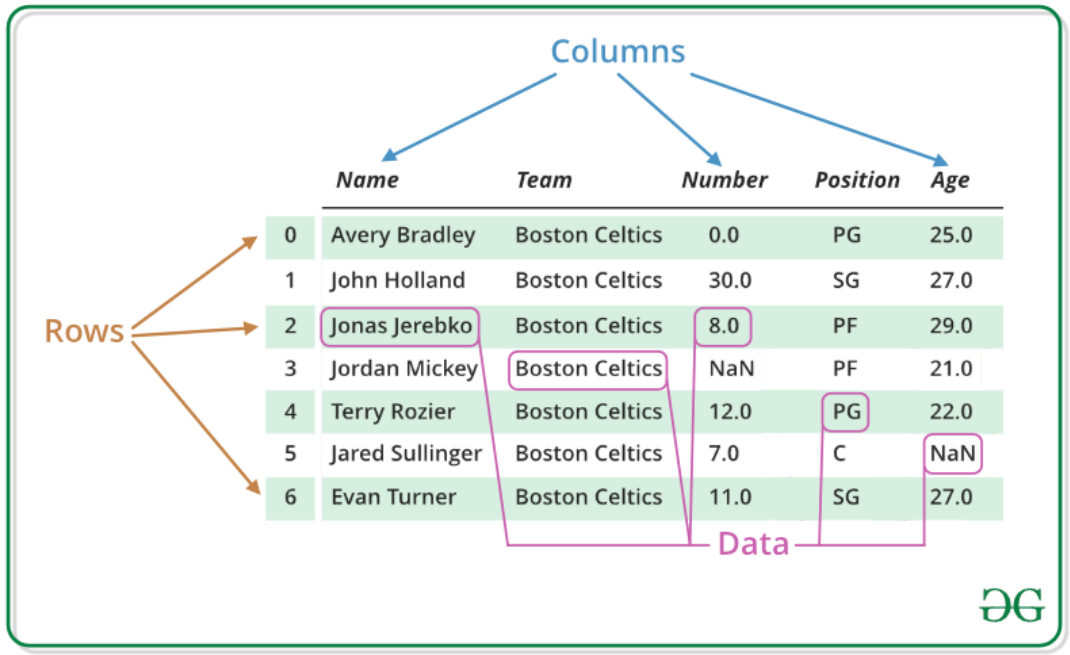
✓ 行：一个样本；

✓ 列：一个字段；

结构化数据特点：

✓ 所有的样本有用相同个数的字段

✓ 每个字段的类型相同；



Part3 入门结构化竞赛

Tableaur dataset

Kaggle中结构化竞赛的分类：

- ✓ 单表单id：所有的记录存储在单张表格内，且样本与标签存在一一对应关系；
- ✓ 单表多id：所有的记录存储在单张表格内，且多条样本记录对应与一个标签；
- ✓ 多表单id：所有的记录存储在多张表格内，且样本与标签存在一一对应关系；
- ✓ 多表多id：所有的记录存储在多张表格内，且多条样本记录对应与一个标签；

Main table in sheet1

	A	B
1	ID	Price
2	1001	2
3	1002	3
4	1003	2
5	1004	4
6	1005	1

New data table in sheet 2

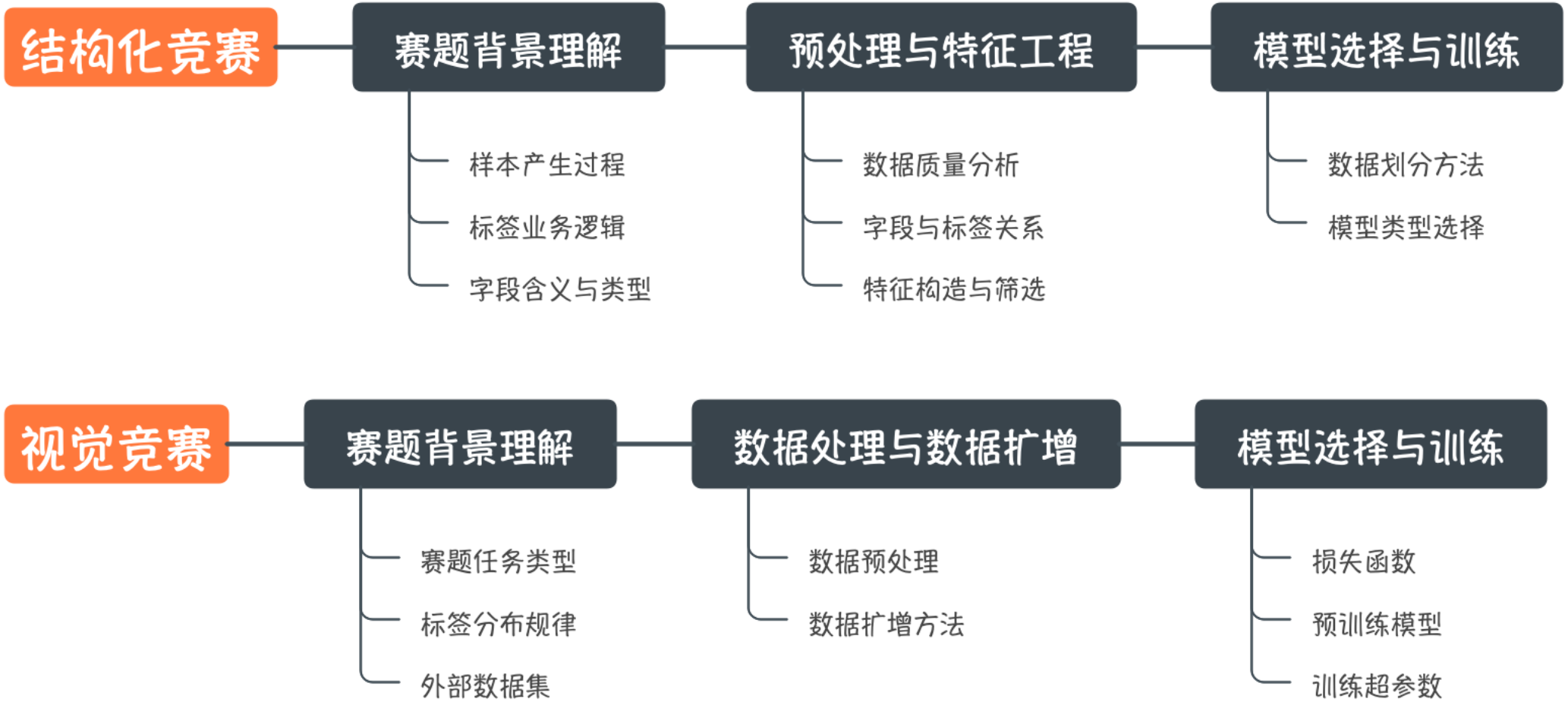
	A	B	C
1	ID	Store	Location
2	1001	A	N
3	1003	A	N
4	1004	C	W
5	1002	B	E
6	1005	C	W

Merge and update in main table

	A	B	C	D
1	ID	Price	Store	Location
2	1001	2	A	N
3	1002	3	B	E
4	1003	2	A	N
5	1004	4	C	W
6	1005	1	C	W

Part3 入门结构化竞赛

Tableaur dataset



Part3 入门结构化竞赛

Tableaur dataset

表格赛	CV赛
任务多样	任务固定
依赖人工	依赖机器和经验
主要是CPU资源	主要是GPU资源

- ✓ 与表格赛相比，CV赛题任务更加固定（分类、分割和检测）；
- ✓ 与表格赛相比，CV赛题需要更多的计算资源，奖金相对多（人少，钱多）；

Part3 入门结构化竞赛

Tableaur dataset

1.代码基础

基础技能包括：Pandas、数据分析技能、树模型使用与调参；

进阶技能包括：模型集成、特征工程

2.关键代码部分

步骤1：对数据集进行数据分析

步骤2：对数据集字段进行编码

步骤3：使用LightGBM进行五折交叉验证

Q&A

Ask me anything

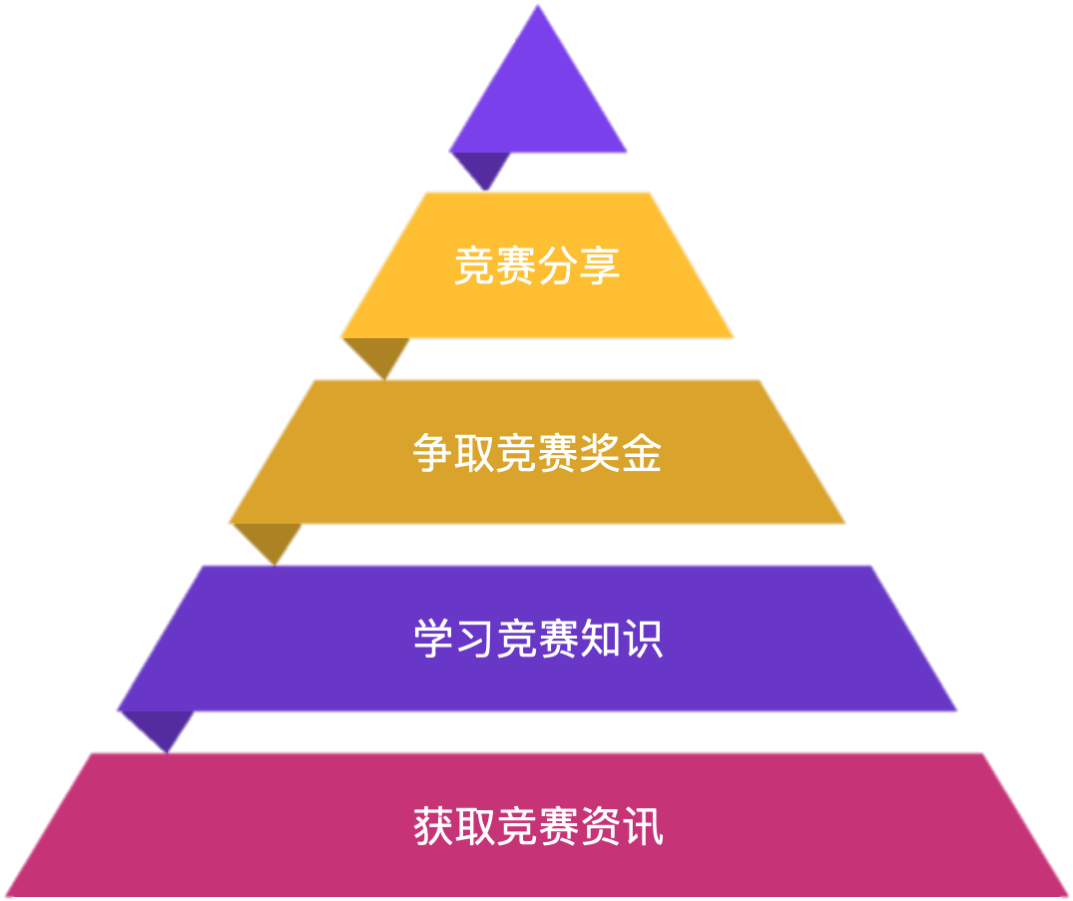
- 竞赛学习路径

✓ 每位同学参与竞赛的目的不同

知识/奖金/认可？

✓ 每位同学参与竞赛的背景不同

本科/找工作/就业？



Q&A

Ask me anything

- 竞赛学习路径

- ✓ 入门阶段（一周）：掌握数据挖掘流程、Pandas、Sklearn
- ✓ 进阶阶段（四周）：掌握特征工程、特征筛选、掌握XGBoot、LightGBM
- ✓ 深入阶段（半年）：
  - ✓ 深度学习基础：Pytorch、Keras、模型搭建、训练流程、深度学习调参
  - ✓ NLP领域知识：TFIDF、Word2Vec、TextCNN、Bert
  - ✓ CV领域知识：CNN、预训练模型、分类模型、检测模型、分割模型