中国建设银行
China Construction Bank

建行杯

人工智能大赛

ARTIFICIAL INTELLIGENCE COMPETITION

2023

题目解析

基于对公信贷审批意见的财务风险判断

01. 赛题介绍

02. 思路分享

03. 经验交流

04. 常见问题汇总

赛题介绍

【赛题背景介绍】

在对公信贷审批环节，专职审批人会基于客户的基本信息、所处行业、历史表现、信贷用途、财务报表等信息，综合评估是否审批通过申请的额度。在审批人的批复意见中，包含其基于企业财务报表数据分析得到的企业经营情况、还款能力及潜在的财务风险等内容。

企业财务风险的判断，一般采用违约相关的数据作为标注数据进行建模分析，但违约数据往往较少，不足以支撑人工智能模型的训练。因此，可以利用额度授信审批环节中，审批人指出的关于企业财务风险的文本信息，对其进行文本挖掘，生成企业财务风险标签，与违约数据相结合，为后续财务报表指标数据评估提供数据基础。

赛题介绍

【赛题任务说明】

本赛题考察选手数据处理能力和对**文本分类问题**的建模能力，要求参赛者利用审批批复的文本数据和标注的财务风险标签，选择合适的算法，建立模型来判断批复意见中是否描述了企业的财务风险。

示例

ID	审批批复意见	标签
1	审批会议提示：鉴于借款人资产负债率偏高。后续.....	1
2	持续条件为流动比率不得低于1。贷后方面,.....	0

赛题介绍

【得分计算说明】

本次比赛采用模型在测试数据集上的F1-score对模型预测性能进行评分，定义TP：样本为正，预测结果为正；FP：样本为负，预测结果为正；TN：样本为负，预测结果为负；FN：样本为正，预测结果为负。

其中，准确率Precision计算公式为：$Precision = \frac{TP}{TP+FP}$

召回率Recall计算公式为：$Recall = \frac{TP}{TP+FN}$

F1-score计算公式为：$F1-score = \frac{2 \times Precision \times Recall}{Precision+Recall}$

最终排名：按 F1-score 从高到低排序。

赛题介绍

【数据说明】

数据分为两部分：训练数据及测试数据，均来源于行内额度授信审批过程中的批复数据，审批意见文本内容中企业名称、业务编号等内容已经过脱敏处理。

训练数据包含1个文件：train.csv。该文件包含三个字段：

- ID：唯一标识；
- content：批复意见文本内容；
- label：是否有财务风险。

训练集为带有财务风险标签值（label）的数据集，其中label=0表示无财务风险描述，为1则表示有财务风险描述。训练集的标签是基于业务提出的一些规则进行标注。

赛题介绍

【数据说明】

测试数据包含1个文件：
测试A榜：testA.csv；
测试B榜：testB.csv，在复赛阶段开放。

测试数据文件字段为ID（唯一标识）、content（批复意见文本内容），无label字段。A榜、B榜两份数据，分别用于初赛和复赛阶段。

赛题介绍

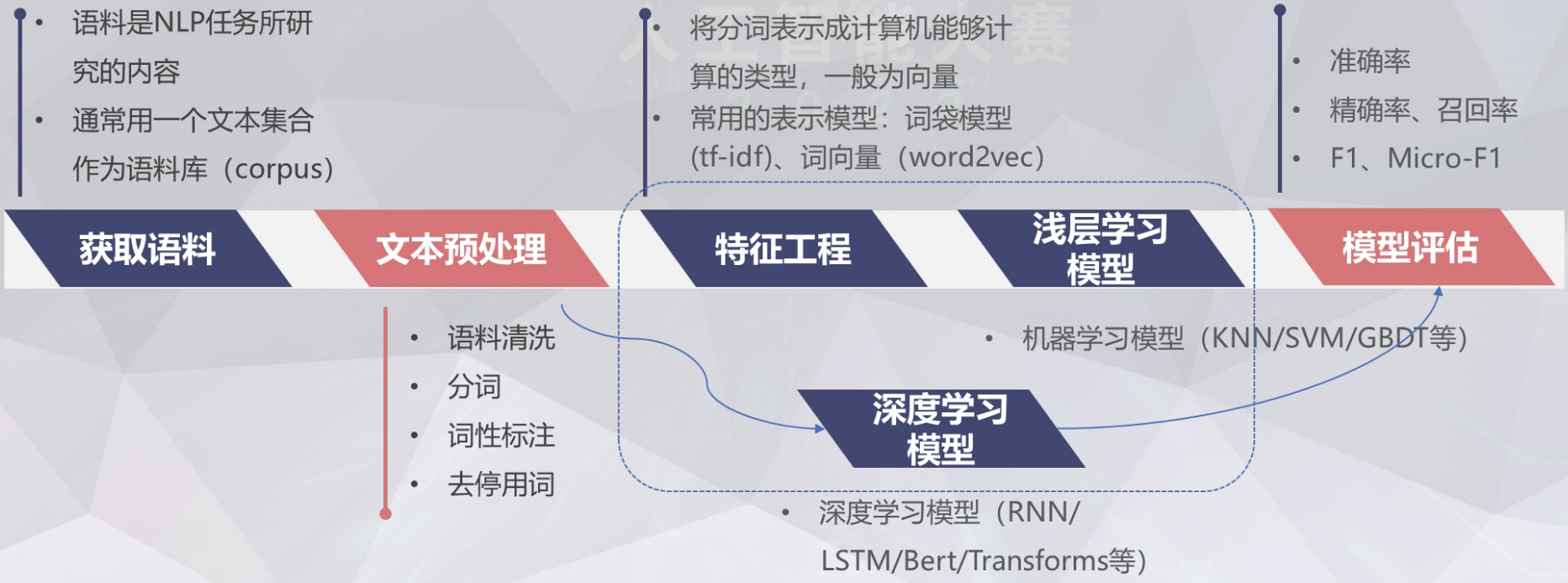
【数据说明】

- 审批批复意见的文本内容长短不一，有的意见文本会比较长，一两千字。审批意见的内容包含的信息比较多，不仅只描述企业的财务，还涉及一些企业本身的基础信息及所属集团信息、一些贷后持续条件、关联企业情况等。
- 本赛题的主要目标是识别企业自身的财务方面风险，有些描述的是关联企业或集团、担保人等的负面评价都不算是正标签。
- 文本数据有对企业名称和相关的编号做了脱敏处理，以***代替。
- 存在数据不均衡问题，正标签数量相对较少。

思路分享

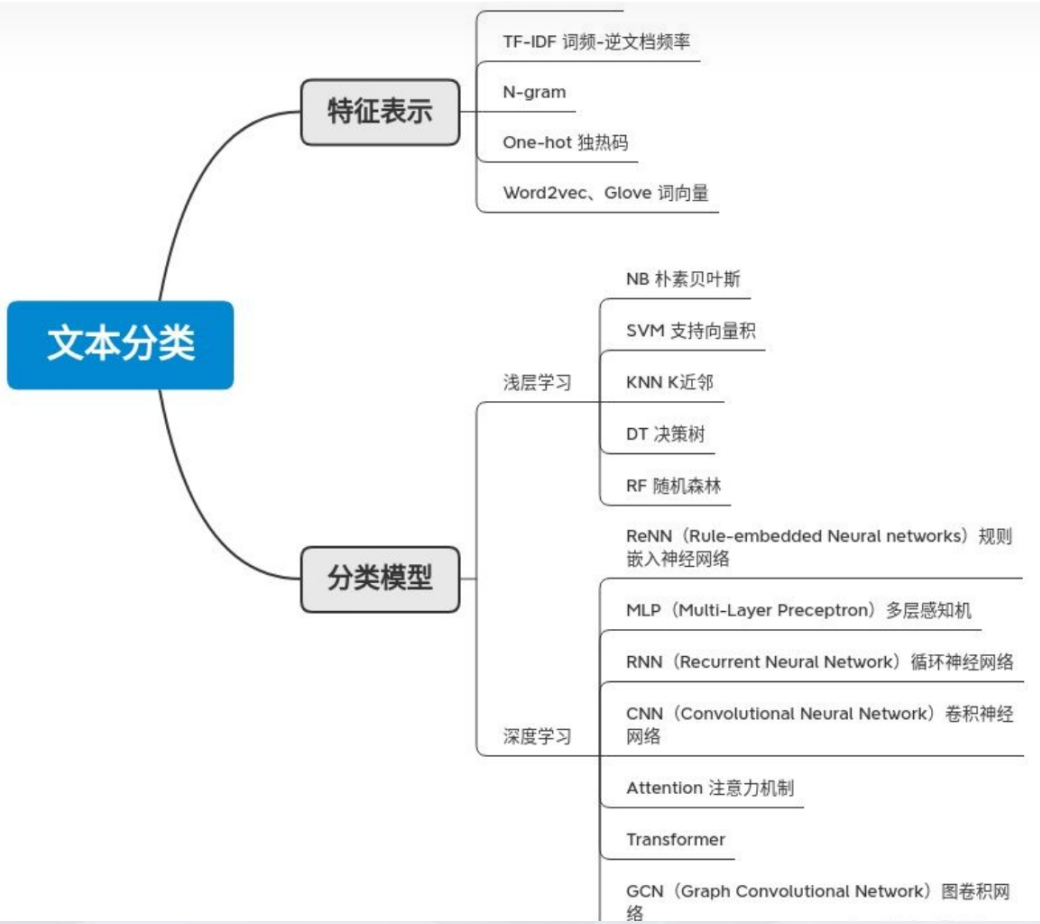
【文本分类建模】

本赛题本质上是一个文本二分类问题。通常文本分类可通过以下几步进行建模，主要步骤是文本处理和算法选择。算法方面，可选择传统的机器学习方式，也可采用深度学习模型，通常情况下，深度学习的效果会比较好。



思路分享

【文本分类建模】



- 浅层学习模型结构较为简单，依赖于人工获取的文本特征，虽然模型参数相对较少，但是在复杂任务中往往能够表现出较好的效果，具有很好的领域适应性。

- 深度学习模型结构相对复杂，不依赖与人工获取的文本特征，可以直接对文本内容进行学习、建模，但是深度学习模型对于数据的依赖性较高，且存在领域适应性不强的问题。

思路分享

【思路一】

词向量（word2vec）+ 机器学习分类算法（SVM/KNN等）
或直接使用fasttext训练文本文本分类模型得到结果，简单的步骤如下：

读取数据 → 文本处理 → 模型训练 → 模型评估

```
train_data = pd.read_excel("data/train.xlsx")
```

```
train_data['content'] = train_data['content'].apply(lambda x:x.replace('\n','。'))
train_data['sents_cut'] = train_data['content'].apply(lambda x:jieba.lcut(x))
train_data['corpus'] = list(map(lambda x,y: '__label__' + str(x) + '\t' + " ".join(y),
                                train_data['label'], train_data['sents_cut']))
train_data['sents_cut_join'] = list(map(lambda x: " ".join(x),train_data['sents_cut']))
```

```
train_dataset, test_dataset = train_test_split(train_data, test_size=0.3, stratify=train_data['label'])
```

```
with open('data/corpora.train','w') as f:
    f.write('\n'.join(train_dataset['corpus'].tolist()))
with open('data/corpora.test','w') as f:
    f.write('\n'.join(test_dataset['corpus'].tolist()))
```

```
st = time.time()
model = fasttext.train_supervised(input='data/corpora.train',label='__label__',dim=100,
                                  wordNgrams=3, lr=0.1, thread = 8, loss='ns',
                                  autotuneValidationFile='data/corpora.test', autotuneDuration=300)
train_result=model.test('data/corpora.train')
print('train_precision:', train_result[1])
print('train_recall:', train_result[2])
print('Number of train examples:', train_result[0])
```

【思路二】

采用bert + finetune的方式。

- 有些样本数据较长，会超过512的字数限制，需对文本进行切分或截断等处理。
- 训练数据不多，可以通过数据增强或算法层面，防止过拟合。

【思路三】

大语言模型+prompt工程

- 可采用few-shot构建prompt。

【思路四】

大语言模型微调

经验分享

文本分类是较常见的NLP任务。本赛题需要注意的点：

- 不同于情感分析，文本内容可能存在其他与财务无关的负面评价，不能单纯使用情感判别模型。
- 有些审批意见很长，可能还涉及这个额度授信方案的描述。如果使用bert相关的模型，需要注意文本的长度，进行文本处理。
- 文本中很多与财务无关的描述，可适当做一些文本的清洗。往往和企业本身财务相关的描述就是一两句，而标注的是整个样本的标签，而没有单句的描述。
- 数据存在不平衡问题，可进行数据增强和模型融合等。

常见问题汇总

【标签的标注标准】

- 针对批复意见中提到的企业本身的财务指标相关风险进行标注，有些只提到需进行关注，无具体财务相关的负面描述，有些描述的只是关联企业、所属集团或担保人的财务风险，这样的样本标签也为0。
- 有些没提到具体的财务指标问题，而笼统地说企业规模小、与授信额度不匹配等，标签也为0。
- 训练数据的标签存在一些噪声数据，由于数据量较多，人工标注不完全准确。

【训练集与测试集】

- 训练集与测试集的正负样本不平衡，有财务风险的描述相对少一些。训练集和测试集来源的时间范围是一致的，但行业信息不一样，有些财务指标的描述会存在差异。

【提交数据的格式】

- 最终上传结果只支持对于0/1分类值的打分。不能提供预测概率值或文本描述。

感谢聆听