基于 WCG 模型分析关于 Kaggle 竞赛的 Titanic 报告

学号:20192131089　20192131077　姓名: 吴宇涛 张帆

2021 年 9 月 22 日

目录

摘要

近年来机器学习成为一个热门学科, 高等院校在计算机专业都有开设机器学习相关课程. 实践参加一次 kaggle 比赛并总结一份报告是快速入门机器学习的方法。本文通过总结机器学习 kaggle 竞赛《Titanic - Machine Learning from Disaster》, 完成对乘客是否幸存的预测, 以及阐述如何认识数据特征值, 如何处理缺失值, 数据可视化, 使用 WCG 模型对 WCG 乘客进行预测.

关键词: Kaggle; Titanic; 机器学习; WCG 模型

1 问题重述

1.1 问题背景

泰坦尼克号的沉没是历史上最臭名昭著的沉船之一。

1912 年 4 月 15 日，在她的处女航期间，被广泛认为“不沉”的泰坦尼克号在与冰山相撞后沉没。不幸的是，船上的每个人都没有足够的救生艇，导致 2224 名乘客和船员中有 1502 人死亡。

虽然幸存下来有一些运气因素，但似乎有些人比其他人更有可能幸存下来。

1.1.1 问题提出

在这个 Kaggle 挑战中，要求构建一个预测模型来回答这个问题：“什么样的人更有可能生存?”使用乘客数据（即姓名、年龄、性别、社会经济阶层等）。

2 数据分析

2.1 获取数据

表 1: df.head()

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
0	1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
2	3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
4	5	0.0	3	Allen, Mr. William Henry	male	35.0	0

从 Kaggle 官网中下载三个.csv 表, 使用 pandas 库读取 train.csv, test.csv, 并且将它们合并为一个 dataframe, 从二维表可以看出我们有 12 列即 12 个特征. 其中 Survived: 因变量, 训练集里面都是 0 或者 1, 表示是否生存, 测试集里面都是 NaN. 我们需要通过数据处理来完成对缺失值的处理, 并且确定哪些特征具有较强的预测能力, 哪些是中等预测能力特征, 哪些是弱预测能力特征.

2.2 数据处理

2.2.1 分析特征值

特征:

- Pclass: 表示阶层, 越小表示阶层越高, 从生存率高低或者 Fare 的平均值高低可以看出这一点.
- Sex: 性别, ” 女士孩子优先” 原则让性别对生存率影响较大.
- Age: 年龄, ” 女士孩子优先” 原则应该会让未成年人的生存率提高, 其他各个年龄段的生存率是否不同呢?
- Fare: 船票的总价 (一张船票有可能让好几个乘客上船)
- Embarked: 上船地点. 粗看起来没有什么用处, 但是, 如果考虑与其他乘客之间的关系的话, 对预测生死有可能有帮助的特征:
- Name: 姓名好像无法决定生死
- Ticket: 船票号码
- Cabin: 船舱号码

- SibSp: 兄弟姐妹或者配偶的个数
- Parch: 父母或者孩子的个数

2.2.2 观察缺失值

- Survived 缺失的都是测试集的部分
- Cabin, Age 缺失比较严重, 如果一定要使用它们, 需要比较谨慎, 后来发现它们并不重要, 没有使用上.
- Embarked, Fare 缺失较少

表 2: 缺失值

Cabin	1014
Survived	418
Age	263
Embarked	2
Fare	1
dtype	int64

2.2.3 数据探索性分析

引入新特征 1: Title

- “女士和孩子优先 “原则下, 不需要特别分离出女孩了, 女孩和成年女性分为一类即可
- “女士和孩子优先 “原则下, 男孩有必要从男性中分离出来, 男孩称为 “boy“, 成年男性称为 “man“
- 为了和原先的特征 “Sex“作区分, 称为 “Title“, 这两个特征的信息量是有比较大的重叠的, 可以考虑保留其中一个.

表 3: 新特征:Title

man	0.597403
woman	0.355997
boy	0.046600
name:Title	dtype: float64

- man 占了 59.7
- 后面会发现这个占比很重要, man 的占比很大, 但难以预测出幸存者, 这就是为什么最高分 (83% 左右) 难以超出基准分 (性别模型, 76.6%) 太多的主要原因.

引入新特征 2: Pfare

- Fare: 船票的总价 (一张船票有可能让好几个乘客上船)
- 因此, Fare 是一个交叉性特征, 同时考虑了平均价格和一张票的乘客数量成年男性称为 “man“
- 对 Fare 这个特征进行” 提纯”, 求平均价格 (Pfare), 平均价格更能体现出乘客的地位, 从而反映在生存率上.
- 有一些 Fare 取值为 0, 这个设定不能接受, ” 没有免费的午餐”, 将 0 赋值为 NaN.

2.3 数据可视化

2.3.1 新特征:Title

- 发现 male 生存率远远低于平均值, female 生存率远远高于平均值, 这就是为什么简单的"gender model", 正确率就高达 76.6 % 的原因.
- "gender model": 预测完全只看 Sex 特征, 如果是 male, 一律预测为死亡, 如果是 female, 一律预测为生存.
- 发现 boy 虽然也是男性, 但是生存率与 man 截然不同, man 生存率不到 20 %, 而 boy 生存率接近 60 %
- 说明对男性分成细分成两类, 是有价值的
- 有遗憾的地方在于: boy 的占比太小了, 对整体成绩的提升非常有限 (理论上, 准确率最多提高 4.7 % 左右)

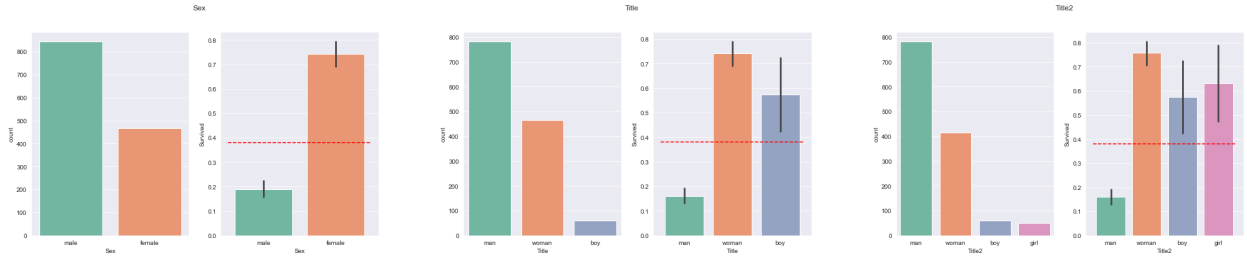

图 1: ‘sex’ 特征和 ‘title’ 特征

2.3.2 特征:Pclass

通过观察表 1 以及图 2 的数据可以发现等级越高, 生存率越高, 这符合常识, 也符合我们的预期.

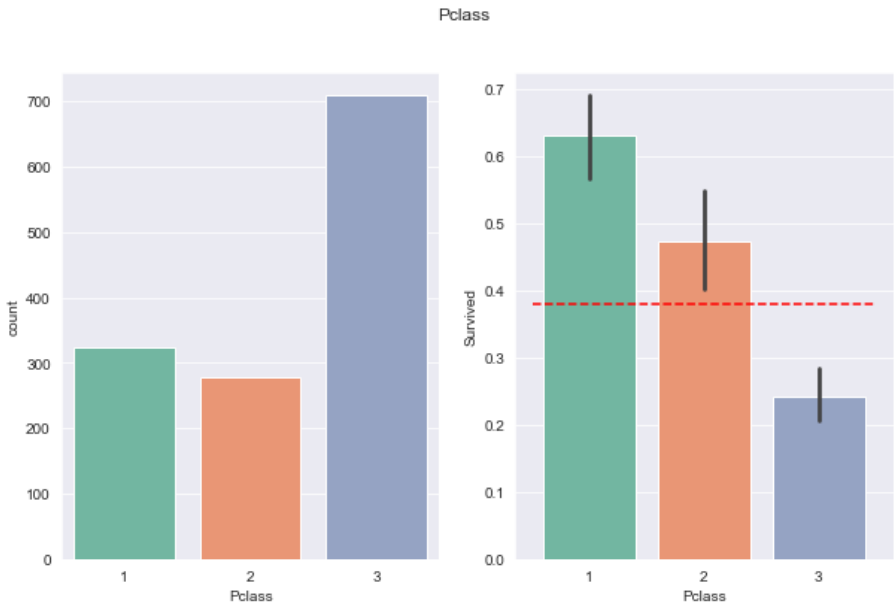

图 2: ‘pclass’ 特征

2.3.3 特征: Embarked

通过观察图 3, 我们可以得知:

- 取值为 C 时, 高于平均值

- 取值为 S 时, 低于平均值
- 取值为 Q 时, 在平均值附近
- 总体来看, 是中等预测能力的特征.

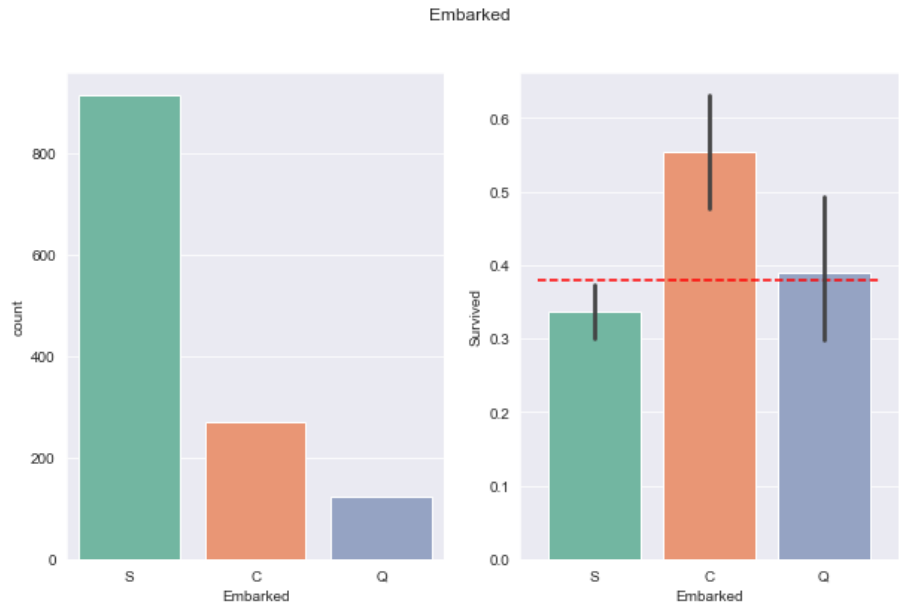

图 3: ‘Embarked’ 特征

2.4 特征工程

基于数据可视化, 我们可以对特征值进行分类, 分别是强, 中, 弱预测能力特征.
强预测能力特征:

- Title, Sex
- Pclass
- Pfare

中预测能力特征:

- Embarked

弱预测能力特征:

- Family, Parch, SibSp

通过分析以上特征值的预测能力, 对于弱预测能力特征:Family, Parch, SibSp, 因为这些都是家庭因素, 所以可以不再考虑. ’Age’ 是无预测能力特征可以直接忽略. Title 是最强的特征, 本文后面将针对 Tittle 特征着重介绍 WCG 模型和’Votting of KNN’ 模型.

删除其他无关特征, 降低特征个数, 方便我们后面使用 WCG 模型进行训练测试.

3 使用 WCG 模型对 WCG 乘客进行预测

3.1 WCG 模型的核心概念

WCG 不同于 WCG 模型, WCG 是一种分组方法, WCG 模型是基于 WCG 的预测模型. WCG 用一句话来概括, 就是: 将一个家庭内的所有女性 (包括女保姆) 或孩子认定为一组. 需要特别留意的是: 一个家庭内的成年男性并没有分到组里面.

	PassengerId	Survived	Pclass	Name	Ticket	Fare	Embarked	Title	Pfare	Pfare_bin
0	1	0.0	3	Braund, Mr. Owen Harris	A/5 21171	7.2500	S	man	7.25000	(7.229, 7.75]
1	2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	PC 17599	71.2833	C	woman	35.64165	(31.679, 128.082]
2	3	1.0	3	Heikkinen, Miss. Laina	STON/O2. 3101282	7.9250	S	woman	7.92500	(7.896, 8.05]
3	4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	113803	53.1000	S	woman	26.55000	(26.283, 31.679]
4	5	0.0	3	Allen, Mr. William Henry	373450	8.0500	S	man	8.05000	(7.896, 8.05]

图 4: 特征降维

一个家庭内的成员活动在一起, 同生或同死的概率极大, 可以利用这个性质, 如果一个组内的成员一部分在训练集里面, 一部分在测试集里面, 根据训练集的生存率就可以去预测测试集成员的生死. 男人因为” 女士和孩子优先” 这条规则, 并不遵循第 1 条逻辑. 虽然传统意义上, 女保姆不算是家庭成员, 但是女保姆和东家活动在一起, 也符合” 女士和孩子优先” 这条规则的, 因此, 女保姆的生存率和东家的女生孩子的生存率应该是没有任何区别的.

3.2 非 WCG 乘客 (“noGroup”) 与 WCG 乘客 (非 “noGroup”) 的区别

通过附件中提供的源代码运行可得如下结果: 只有占比 230/1309 的乘客是 WCG 乘客, 另外占比 1079/1309 的乘客需要其他模型来进行预测. 大约有 96.7 % 的 boy 是 WCG 成员. 大约有 36.7 % 的 woman 是 WCG 成员. 没有 man 是 WCG 成员.

表 4: 非 WGC 乘客和 WCG 乘客的却别

非 WCG 组号个数:	1
WCG 组号个数: 80	80
非 WCG 乘客个数: 1079	1079
WCG 乘客个数: 230	230
非 WCG 乘客的 Title 分布:	
man	782
woman	295
boy	2
Name: Title, dtype: int64	
WCG 乘客的 Title 分布:	
woman	171
boy	59
Name: Title, dtype: int64	
全体乘客的 Title 分布:	
man	782
woman	466
boy	61
Name: Title, dtype: int64	
非 WCG 覆盖率:	
man	1.000000
woman	0.633047
boy	0.032787
Name: Title, dtype: float64	
WCG 覆盖率:	
boy	0.967213
man	0.000000
woman	0.366953
Name: Title, dtype: float64	

3.3 WCG 模型成绩评估

WCG 多增加了 21 个正确预测模型, 但是相比 “gender model” 只改变了 23 个样本 (其中 8 个 boy, 15 个 woman) 这意味着这 23 个样本里面, 有 22 个正确预测, 1 个错误预测, 因为将原先正确的结果改成了错误的结果, 要倒扣分. 这 23 个样本里面, 有 22 个正确, 正确率奇高!

4 预测与提交

4.1 man 的集成模型

表 5: man 的集成模型

	Pfare	Pclass_1	Pclass_2	Pclass_3	Embarked_C	Embarked_Q	Embarked_S
919	1.325405	1	0	0	0	0	1
930	-0.590534	0	0	1	0	0	1
941	1.284533	1	0	0	0	0	1
985	0.957553	1	0	0	1	0	0
1096	0.936438	1	0	0	1	0	0

4.2 woman 的集成模型

表 6: woman 的集成模型

Pfare	Pclass_1	Pclass_2	Pclass_3	Embarked_C	Embarked_Q	Embarked_S	
1060	-0.619789	0	0	1	0	0	1
1088	-0.689270	0	0	1	0	0	1
1105	-0.689270	0	0	1	0	0	1
1250	-0.689270	0	0	1	0	0	1
1267	-0.637342	0	0	1	0	0	1
1303	-0.689270	0	0	1	0	0	1

4.3 输出.csv 文件

```python
df.loc[X_test_man.loc[y_test_hat_man == 1].index, "Predict"] = y_test_hat_man[y_test_hat_man == 1]
df.loc[X_test_woman.loc[y_test_hat_woman == 0].index, "Predict"] = y_test_hat_woman[y_test_hat_woman == 0]

output = pd.DataFrame(
    {
        "PassengerId": df[n_train:].PassengerId,
        "Survived": df[n_train:].Predict.astype("int"),
    }
)
output.to_csv("WCG_non-WCG.csv", index=False)
```

Python

图 5: 运行代码

4.4 分析结果

准确率为 0.82057, 意味着总共有 342 个样本被正确预测, 相比 WCG 模型, 就只多了 1 个样本. non-WCG 模型只改变了 11 个样本, 意味着其中有 6 个改对了, 5 个被改错了, 净收益为 1 个样本. non-WCG 模型的表现不尽如意, 但还是将 non-WCG 模型保留下来, 目的是为了方便以后套用机器学习的模板.

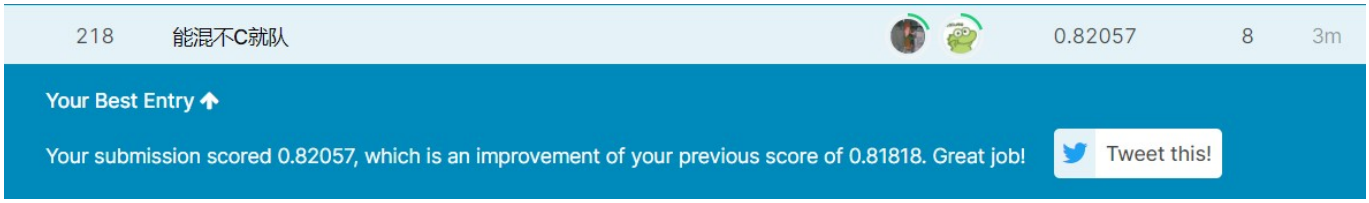

图 6: Kaggle 得分

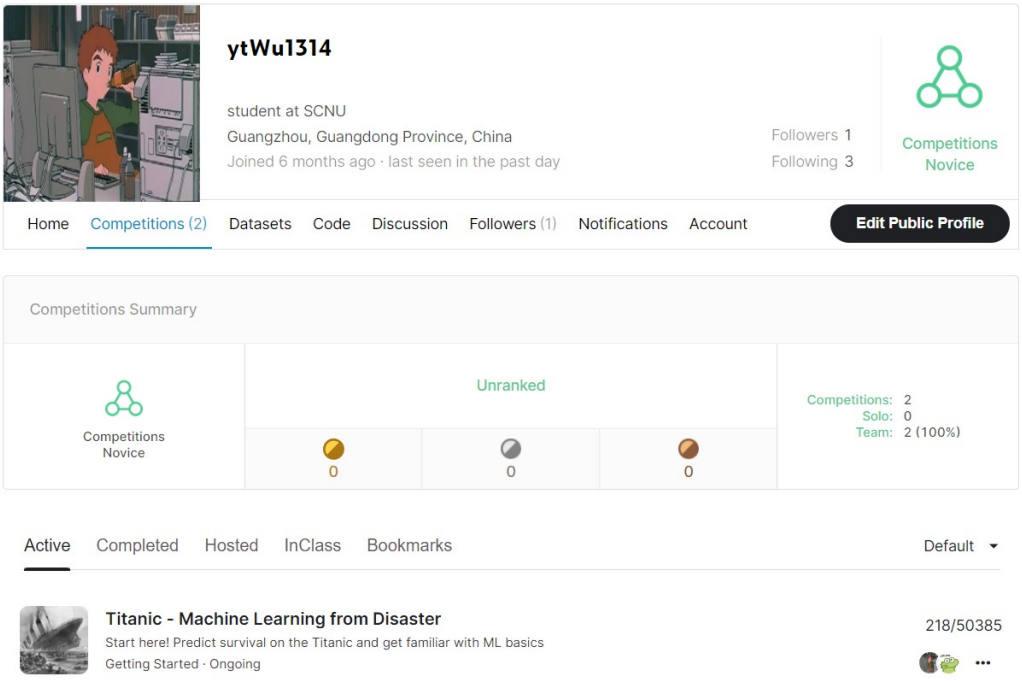

图 7: 成绩排名

5 总结

通过参加本次的 Kaggle 竞赛以及书写项目报告, 基本了解了机器学习的流程, 首先拿到一个数据集, 需要分析哪个是目标值, 哪些是有用的特征值, 对无关的特征要进行降维处理. 需要了解如何调用 sklearn 中的 API.

本次竞赛, 我们使用了 WCG 模型对特征值再进行了处理, 提高了测试的准确率. 使用了集成模型, 对于 man 使用了 KNN, SVC, LGBM. 对于 womean 使用了 KNN, SVC. 基本入门机器学习, 但还是需要后面紧跟课程的学习, 将机器学习中学到的知识灵活运用, 以便准备后面深度学习的大项目.

参考文献

[1] https://github.com/li1127217ye/Titanic

[2] https://nbviewer.jupyter.org/github/li1127217ye/Titanic/blob/main/Titanic_ WCG_ non-WCG.ipynb

[3]《从机器学习到深度学习基于 scikit-learn 与 TensorFlow 的高效开发实战》