机器学习

泰坦尼克号预测算法

主　研　人：董乐诗
参　研　人：董乐诗

审　核　人：
方　　向：工业算法案例研究
版　本　号：A

大数据及人工智能部®

中冶赛迪信息技术有限公司

二〇一八年九月

版本更新

日期	版本	更新描述	作者
2019/4/22	A	初稿	董乐诗

选题表格

时间	竞赛名	竞赛背景描述（50 字以内）	类型（分类/回归）
2019/4/22	Titanic: Machine Learning from Disaster	通过所给数据判断泰坦尼克号乘客是否会遇难	分类

目录

1. 背景描述……4
1.1 竞赛赛题描述……4
1.2 评估指标描述……4
2. 数据来源及描述性统计分析……4
2.1 大赛数据来源……4
2.2 数据的描述性统计……5
2.2.1 数据基本情况描述：……5
2.2.2 数据字段介绍：……5
2.2.3 数据描述性统计……6
3. 优秀算法思路……10
3.1 方案一……10
3.1.1 方案一数据预处理及特征工程部分方案……10
3.1.2 方案一模型设计、建立部分方案……10
3.1.3 方案一结果、排名等……12
3.1.4 方案一算法流程图……12
4. 算法比较……13
表 4-1 算法比较……13
5. 总结与展望……13
5.1 总结……13
5.2 建模思路……13

1. 背景描述

RMS 泰坦尼克号沉没是历史上最臭名昭着的沉船之一。 1912 年 4 月 15 日，在她的处女航中，泰坦尼克号在与冰山相撞后沉没，在 2224 名乘客和机组人员中造成 1502 人死亡。 这场耸人听闻的悲剧震惊了国际社会，并为船舶制定了更好的安全规定。造成海难失事的原因之一是乘客和机组人员没有足够的救生艇。 尽管幸存下沉有一些运气因素，但有些人比其他人更容易生存，例如妇女，儿童和上流社会。

1.1 竞赛赛题描述

在这个挑战中，参赛者需要完成对乘客可能存活的分析。并对于乘客是否可以幸免于难做出二分类判断。

1.2 评估指标描述

提交的结果将用预测结果的正确率百分比来衡量。

提交的结果文件应该有两列：

- PassengerId（按任意顺序排序）
- 生存情况（包含你的二元预测：1 为幸存，0 为死者）

标准格式示例如下：

```
PassengerId,Survived
 892,0
 893,1
 894,0
 Etc.
```

2. 数据来源及描述性统计分析

2.1 大赛数据来源

数据来源：本次大赛的数据来源于 Kaggle 网站

本次比赛官方提供的数据如下：

Titanic

官方数据下载地址如下：

https://www.kaggle.com/c/titanic/data

2.2 数据的描述性统计

2.2.1 数据基本情况描述:

数据被分为了两组：

- 训练集（train.csv）：包含乘客性别、舱等等“特征”
- 测试集（test.csv）：包含和训练集相同的性别、舱等等“特征”
- gender_submission.csv，这是一组假设所有且仅有女性乘客生存的预测，作为提交文件应该是什么样子的示例

2.2.2 数据字段介绍

数据表中的特征描述如下所示：

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

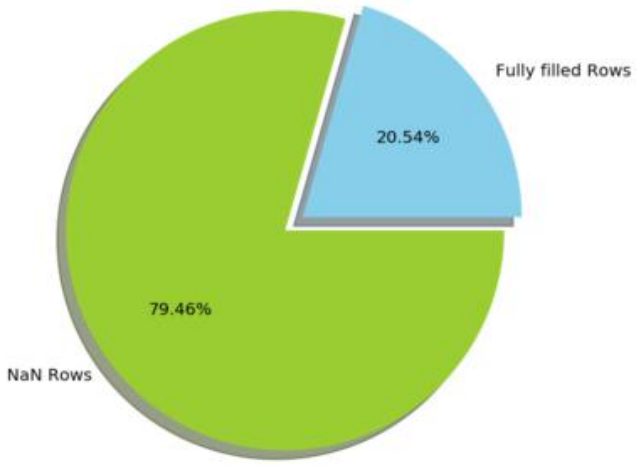
- Survival: 0 = No, 1 = Yes
- pclass: Ticket class 1 = 1st, 2 = 2nd, 3 = 3rd
- sex: Male/Female
- Age: Age in years
- sibsp# of siblings / spouses aboard the Titanic
- parch# of parents / children aboard the Titanic
- ticket: Ticket number
- fare: Passenger fare
- cabin: Cabin number
- embarked Port of Embarkation: C = Cherbourg, Q = Queenstown, S = Southampton

2.2.3 数据描述性统计

①数据集中的数据信息如下所示：

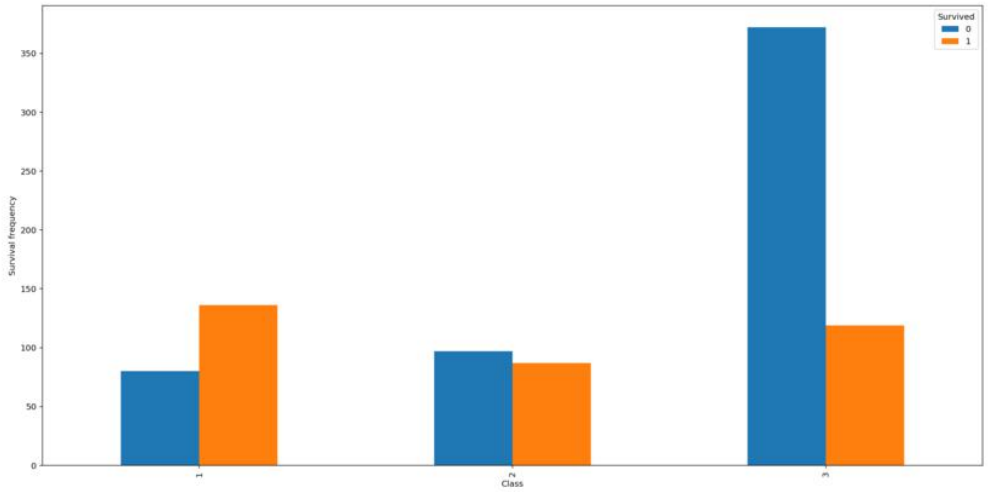
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

②包含 NaN / Null 值的行数如下所示：

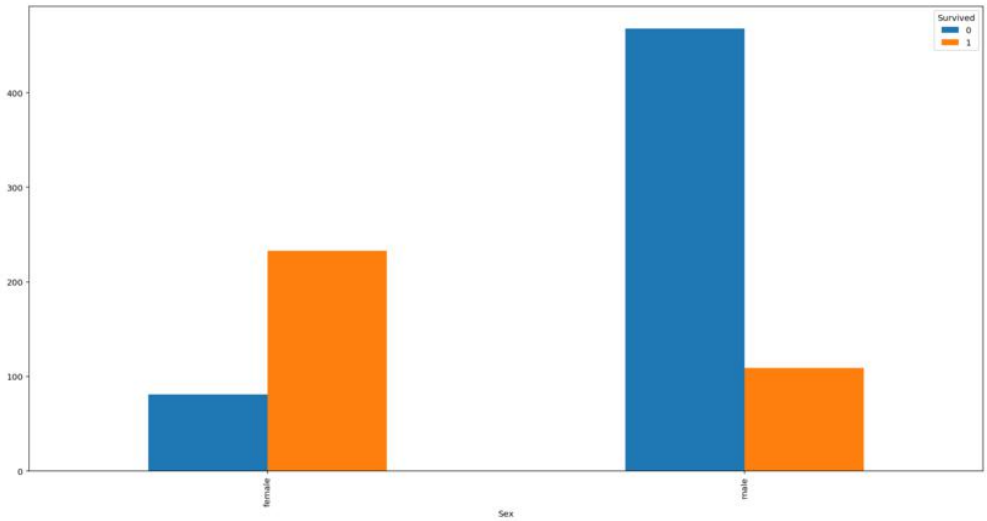


③对于各个特征的数据统计分析如下所示：

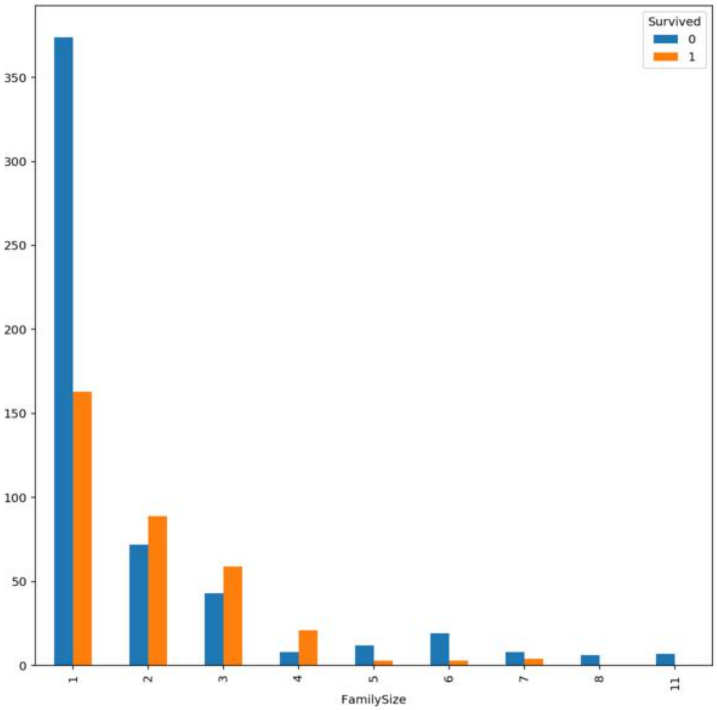
(1) Pclass



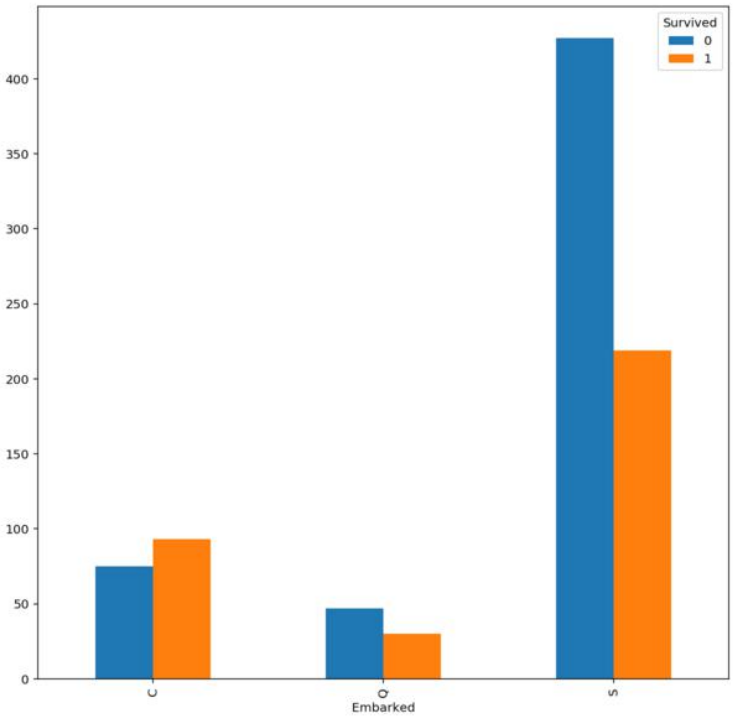
(2) Sex



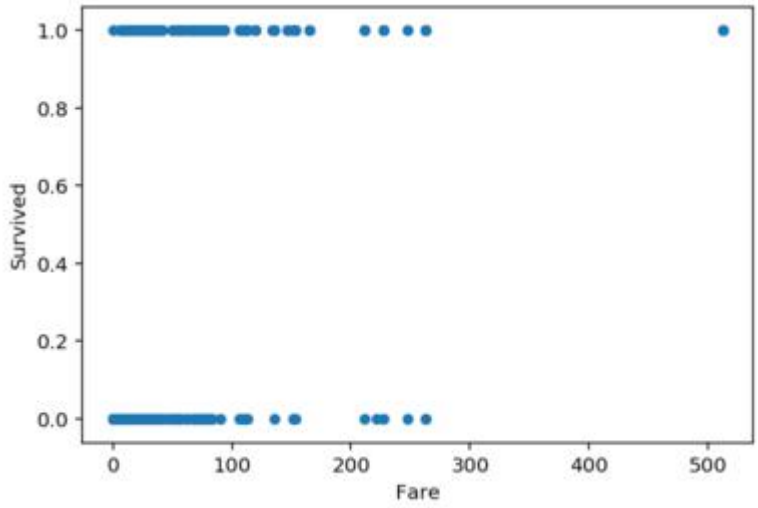
(3) Total Family



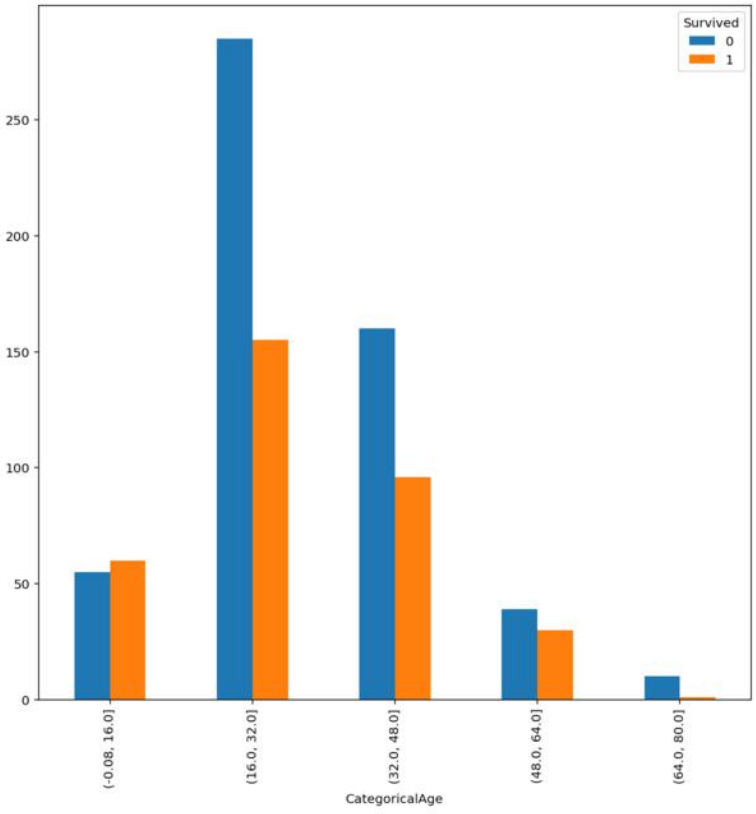
(4) Embarked

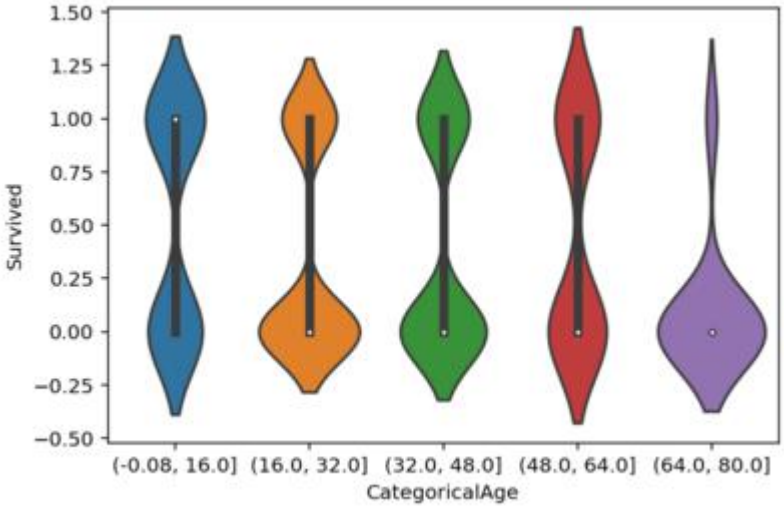


(5) Fare


1.0
0.8
0.6
0.4
0.2
0.0
Survived
0
100
200
300
400
500
Fare

(6)Age


Survived
0
1
250
200
150
100
50
0
(-0.08, 16.0]
(16.0, 32.0]
(32.0, 48.0]
(48.0, 64.0]
(64.0, 80.0]
CategoricalAge



更多详细信息请参阅：https://www.kaggle.com/shriramganesh/diving-deep-into-titanic-data

3. 优秀算法思路

3.1 方案一

由于本项比赛的长期开放性，排行榜上没有固定的 top10 的方案，所以方案一整理的建模思路是一位最近期排名 3%的方案。

3.1.1 方案一数据预处理及特征工程部分方案

①探索性数据分析

用常用的方式观察了数据集的一些统计数据

②处理缺失的年龄值

表现良好的一种简单的方法是用 Pclass 的平均年龄来填补缺失的年龄值

③特征缩放

特征缩放是一种用于规范化独立变量或数据特征范围的方法。对 Fare 变量执行了此操作，用以修正改变量的偏态分布。

3.1.2 方案一模型设计、建立部分方案

①拆分数据集

如下所示：

```
#First we have to split out dataset into X and Y. X being the all the variable and Y being Survi
ved or not i.e. 0 or 1.
x = train.drop('Survived',axis=1)
y = train['Survived']
```

```
#Next we split the dataset into the train and test set
#Test will be 30% of the data and the train will be 70%. By setting test_size = .3
#This way we can test our models predictions on the test set to see how we did.
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.3,random_state=1)
```

②超参数调整

超参数调整是调整算法参数以优化其性能的过程。超参数由数据科学家在训练之前设定。 在参赛选手选择在这里做的随机森林模型的情况下，一些参数包括，估计量的数量，最大深度，最小样本叶子，最小样本分裂等等。这些参数是拆分节点时每棵树考虑的变量。在某些方面它是一个试错过程，如下所示：

```
#from sklearn.model_selection import GridSearchCV
#from sklearn.ensemble import RandomForestClassifier

#rf = RandomForestClassifier(max_features='auto', oob_score=True, random_state=1, n_jobs=-1)

#param_grid = { "criterion" : ["gini", "entropy"], "min_samples_leaf" : [1,2,3,5], "min_samples_
split" : [10,11,12,13], "n_estimators": [350, 400, 450, 500,550], "max_depth":[6,7,8,9]}

#gs = GridSearchCV(estimator=rf, param_grid=param_grid, scoring='accuracy', cv=3, n_jobs=-1)

#gs = gs.fit(train.iloc[:, 1:], train.iloc[:, 0])

#print(gs.best_score_)
#print(gs.best_params_)
#print(gs.scorer_)

#Example of the output
#0.8451178451178452
#{'criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 1, 'min_samples_split': 11, 'n_estima
tors': 375}
```

③随机森林模型的建立

具体代码如下所示：

```
#Building the Random Forest Classification model
from sklearn.ensemble import RandomForestClassifier
rfmodel = RandomForestClassifier(random_state=0,n_estimators=450,criterion='gini',n_jobs=-1,ma
x_depth = 8,min_samples_leaf=1,min_samples_split= 11)
#Fitting the model to x_train and y_train
rfmodel.fit(x_train,y_train)
#Predicting the model on the x_test
predictions = rfmodel.predict(x_test)
```

更多经验总结和方法概括，详情请参考：https://www.kaggle.com/danielv7/basic-approach-for-top-3-in-the-titanic

3.1.3 方案一结果、排名等

结果：得分 0.83253

排名：最终排名 3%

3.1.4 方案一算法流程图

4. 算法比较

表 4-1 算法比较

	评估指标数值	特征工程	基础算法	基本库	语言
算法一	0.83253	特征缩放	随机森林	Numpy、pandas 等	Python

5. 总结与展望

5.1 总结

本赛题是机器学习较简单的分类问题，特征的处理是建模过程最关键的部分。

5.2 建模思路

方案一参赛选手解题思路的流程图如下所示：