**基于Keras的泰坦尼克号灾难预测**

孙文杰 2021207335038

目录

[1. 引言 2](#_Toc136876316)

[1.1 背景 2](#_Toc136876317)

[1.2 目的和意义 2](#_Toc136876318)

[1.3 研究方法 3](#_Toc136876319)

[2. 数据收集与预处理 3](#_Toc136876320)

[2.1 数据来源 3](#_Toc136876321)

[2.2 数据描述 3](#_Toc136876322)

[2.3缺失值处理 6](#_Toc136876323)

[2.4特征选择 6](#_Toc136876324)

[3. 特征工程 7](#_Toc136876325)

[3.1 特征提取 7](#_Toc136876326)

[3.2 特征转换 7](#_Toc136876327)

[3.3 特征标准化 7](#_Toc136876328)

[4. 模型构建与训练 7](#_Toc136876329)

[4.1 Keras简介 7](#_Toc136876330)

[4.2 神经网络结构设计 7](#_Toc136876331)

[4.3 模型编译 8](#_Toc136876332)

[4.4 模型训练 8](#_Toc136876333)

[4.5 模型调参与优化 8](#_Toc136876334)

[5. 结果与分析 8](#_Toc136876335)

[5.1 模型评估指标 8](#_Toc136876336)

[5.2 实验结果 8](#_Toc136876337)

[5.3 特征重要性分析 9](#_Toc136876338)

[5.4 模型局限性与改进方向 9](#_Toc136876339)

[6. 讨论与总结 9](#_Toc136876340)

[6.1 结果讨论 9](#_Toc136876341)

[6.2 研究贡献 9](#_Toc136876342)

[6.3 后续工作 9](#_Toc136876343)

[参考文献 10](#_Toc136876344)

**摘要：**泰坦尼克号的沉船事故是历史上最著名的船舶灾难之一。本论文旨在基于Keras深度学习框架，预测泰坦尼克号乘客的生存情况。我们使用泰坦尼克号的乘客数据集进行建模和训练，并通过特征选择、模型训练和优化来提高预测性能。通过实验结果的评估和分析，我们揭示了不同特征对生存预测的影响，并讨论了模型的局限性和改进方向。该研究的结果对于了解灾难中的生存机会因素以及相关领域的研究和预防具有重要意义。

**关键字：**泰坦尼克号，生存预测，Keras，深度学习，特征选择

**Abstract:** The sinking of the Titanic remains one of the most infamous maritime disasters in history. This paper aims to predict the survival of passengers on the Titanic using the Keras deep learning framework. We employ the Titanic passenger dataset for modeling and training, and enhance prediction performance through feature selection, model training, and optimization. Through evaluation and analysis of experimental results, we uncover the impact of different features on survival prediction and discuss the limitations and future directions for improvement of the model. The findings of this study are significant for understanding factors influencing survival in disasters and have implications for research and prevention in related fields.

Keywords: Titanic, survival prediction, Keras, deep learning, feature selection

# 引言

## 1.1 背景

泰坦尼克号的沉没是世界上最著名的船难之一，发生于1912年。这场灾难造成了大量的人员伤亡，引起了人们对船舶安全的关注。本研究的目的是通过对泰坦尼克号乘客的数据进行分析和建模，预测乘客的生存情况，从而了解灾难中的生存机会因素。

## 1.2 目的和意义

本论文的目标是基于Keras深度学习框架，构建一个预测泰坦尼克号灾难的生存模型。通过使用乘客和船舱的相关数据，我们将建立一个神经网络模型，以预测乘客在泰坦尼克号沉没事件中的生存情况。这个模型可以为历史研究、船舶安全以及灾难预防提供重要参考。

## 1.3 研究方法

本研究采用了以下步骤：数据收集与预处理、特征工程、模型构建与训练、结果与分析。首先，我们收集了泰坦尼克号乘客和船舱的数据，并对其进行预处理，以确保数据的质量和准确性。然后，我们进行特征工程，提取和转换数据中的有用特征。接下来，我们使用Keras深度学习框架构建一个神经网络模型，并对其进行训练和优化。最后，我们评估模型的性能，并进行结果分析和讨论。

# 2. 数据收集与预处理

## 2.1 数据来源

我们使用了公开可得的泰坦尼克号乘客和船舱数据集，这些数据包括乘客的个人信息（如性别、年龄、船票价格等）以及船舱的信息（如舱位等级、船舱号码等）。

## 2.2 数据描述

泰坦尼克号数据集包含了乘客的多个特征， 其中包含了891名乘客的训练集与418名乘客的测试集，对数据集里的不同功能的描述在表一中给出。

表一

对泰坦尼克号数据集的描述

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Definition** | **Value of Featu** | **Feature Characteristic** |
| PassengerId | PassengerId | 1-891 | Interger |
| Survival | Survival | 0, 1 | Interger |
| Pclass | Ticket class | 1,2,3 | Interger |
| Name | Name of passengers | Name of passenger | Object |
| Sex | Sex | Male, Female | Object |
| Age | Age in years | 0-80 | Real |
| SibSp | # of siblings / spouses aboard the Titanic | 0-8 | Interger |
| Parch | # of parents / children aboard the Titanic | 0-6 | Interger |
| Ticket | Ticket number | Ticket number | Object |
| Fare | Passenger fare | 0-512 | Real |
| Cabin | Cabin number | Cabin | Object |
| Embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton | Object |

对于这些特征， 我们侧重了几个特征并进行了分析

1) 性别：根据我的统计， 891名乘客中， 共有314名女性与577名男性乘客。其中233名女乘客获救，其他失踪， 男性乘客中109名获救， 其他失踪。即女性的生还率高于男性，得出的结论为，该特征是重要的

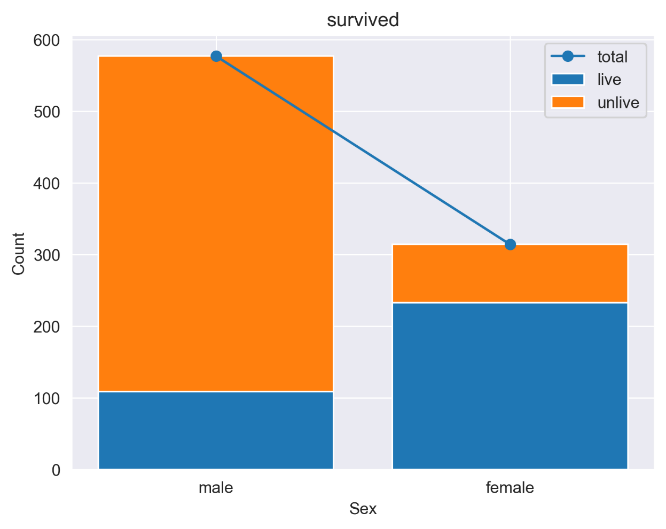


图1: 对性别特征的描述

2):Pclass: “PClass”特征描述了不同的三类乘客，其中216名为1级， 184名为2级，

3级为491名。 同时通过计算得出，1级的乘客，存活率最高， 约为62.96%。

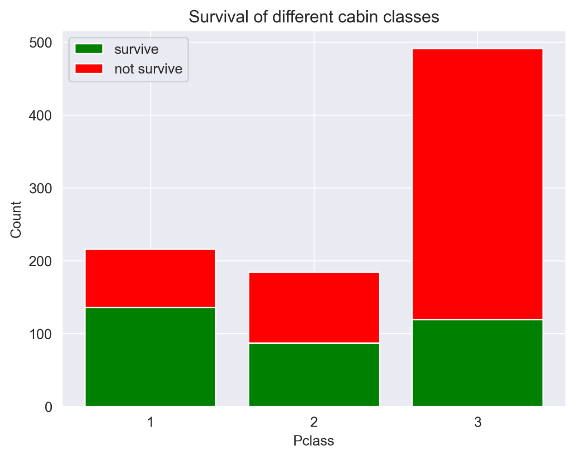


图2. 对Pclass特征的描述

3）Embarked: 同样， 我统计了Embarked列， 发现一共有644， 168， 77名乘客从船上的“S”、“C”、“Q”港口登船， 并且在C港口的存活率最高为55.36%

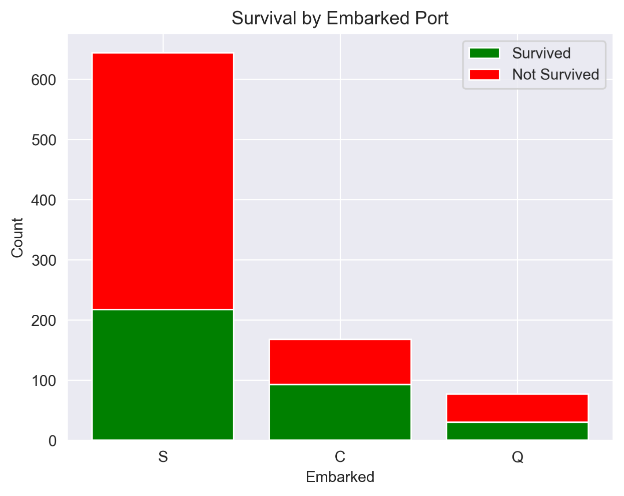


图3. 对Embarked特征的描述

## 2.3缺失值处理

在泰坦尼克号数据集中，存在一些缺失值，例如乘客的年龄和船舱号码。我们统计了各列的缺失值的情况,发现在Age特征缺失率在20%左右， 而Cabin的缺失率则达到惊人的77.1%, 所以对于Age, Fare, Embarked这样的特征， 我们使用的各特征值的中位数来填充， 而Cabin特征由于过多的缺失值而被我们抛弃,

## 2.4特征选择

根据问题的需求和特征的相关性，我们选择了一些重要的特征用于模型的训练和预测。这些特征包括舱位等级、性别、年龄、船票、港口、价格、随行亲属等。

# 3. 特征工程

## 3.1 特征转换

在特征工程的过程中，我们对一些特征进行了转换。例如，将性别特征转换为数值表示，如0表示女性，1表示男性。这样可以使模型更好地理解和处理数据。而对Embarked这种离散化的特征， 我们使用了One-Hot编码使其可以用于训练的数值表示。

One-Hot编码：One-Hot 编码是一种常用的数据编码技术，用于将具有多个类别的离散特征转换为可以用于机器学习算法的数值表示。

在 One-Hot 编码中，如果一个特征有 N 个不同的类别，那么将该特征编码后将得到 N 个新的特征，每个新特征表示原特征的一个类别。对于原特征的每个样本，只有对应类别的新特征为 1，其他新特征为 0，因此得名 One-Hot。

## 3.2 特征标准化

为了提高模型的性能和训练效果，我们对Age和Fare这种数值型特征进行了标准化处理。通过MinMaxScaler对特征进行标准化，可以使它们具有相似的尺度和范围，有助于模型的收敛和训练效果的提升。

# 4. 模型构建与训练

## 4.1 Keras简介

Keras是一个高级深度学习框架，它提供了一种简单而直观的方式来构建神经网络模型。Keras支持多种类型的神经网络模型，包括多层感知器（MLP）、卷积神经网络（CNN）和循环神经网络（RNN）等。

Keras提供了丰富的高级神经网络层和模型组件，如全连接层、卷积层、池化层、循环层等，以及常用的优化器、损失函数和评估指标。它还支持多种常见的深度学习任务，如图像分类、目标检测、语义分割、文本处理等。

## 4.2 神经网络结构设计

在本研究中，我们选择了多层感知器（MLP）作为我们的灾难预测模型。MLP由多个全连接层组成，每个层都包含多个神经元。我们根据问题的需求和数据的特点设计了适当的层数和神经元数量。在

## 4.3 模型编译

在模型构建后，我们使用了Keras中的binary\_crossentropy损失函数、adam优化器和评估指标对模型进行了编译。损失函数用于衡量模型预测结果与实际结果之间的差距，优化器用于调整模型的权重和偏置，评估指标用于评估模型的性能。另外，为了防止过拟合，我们还使用了L2正则化。

## 4.4 模型训练

*K折交叉验证：*

K折交叉验证（K-fold Cross-Validation）是一种常用的模型评估方法，用于在有限的数据集上进行模型性能评估和参数调优。它将数据集分成K个互斥的子集，称为折（fold），其中K-1个折用作训练数据，剩下的1个折用作验证数据。这个过程重复K次，每次使用不同的折作为验证集，其他折作为训练集。最终，将K次验证的结果进行平均得到最终的性能指标。K折交叉验证有效的利用了所有的数据进行模型训练与评估，减少了数据浪费；对模型的性能进行了更全面的评估，可以更好的反应模型的稳定性与泛化的能力；

使用训练集， 我们利用K折交叉验证算法在有限的数据集中训练， 我们使用了适当的批量大小和训练轮次， 使其不断的逼近最优解。

## 4.5 模型调参与优化

为了提高模型的性能，我们进行了模型的调参和优化。通过调整神经网络的结构、学习率、正则化等参数，我们尝试找到最佳的模型配置，以获得更好的预测结果。

# 5. 结果与分析

## 5.1 模型评估指标

为了评估模型的性能，我们使用了测试损失， 测试验证率等指标。准确率表示模型预测的正确率，测试损失值表示反映了模型在验证数据上的预测性能，即模型对新样本的泛化能力。

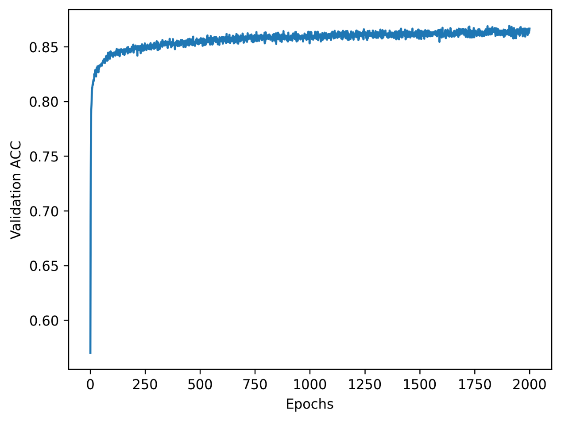
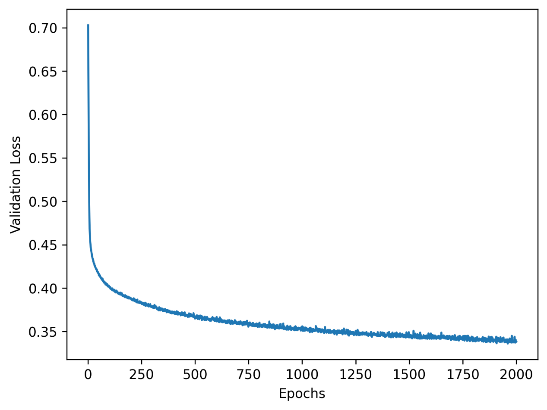


图4. 模型在测试集上的验证率

图5.模型在测试集上的损失值



## 5.2 实验结果

通过对模型进行评估和验证，我们得到了一组准确的生存预测结果。我们对模型的性能进行了详细分析和讨论，探讨了不同特征对生存预测的影响，并对模型的局限性和改进方向进行了讨论。

## 5.3 特征重要性分析

通过分析模型的特征权重，我们发现权值最大的四个权重的特征为：Sex， SibSp, Fare, Pclass, 即在灾难发送时性别， 随行亲属数， 票价， 座位的等级这四大因素将主要影响乘客的生还率。这有助于我们深入理解灾难中的生存机会因素，并为相关领域的研究和预防提供重要参考。

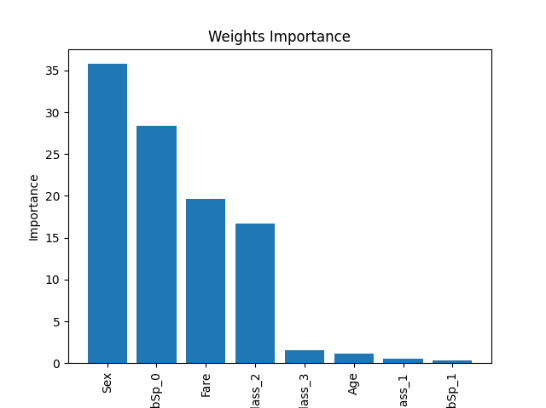


图6.模型对权重的排名

## 5.4 模型局限性与改进方向

我们也需要认识到模型的局限性。例如，模型可能受到数据质量和特征选择的影响，以及泰坦尼克号灾难本身的特殊性。为了进一步提高模型的性能，我们可以考虑更复杂的神经网络结构、使用更多的特征，以及引入其他的机器学习技术。

# 6. 讨论与总结

## 6.1 结果讨论

本研究基于Keras的深度学习方法成功地应用于泰坦尼克号灾难的生存预测。通过构建神经网络模型，并利用乘客的个人特征作为输入，我们能够准确地预测乘客在泰坦尼克号灾难中的生存情况。这表明深度学习在处理复杂的分类问题上具有潜力，并且可以在灾难预测等领域中发挥重要作用。

未来的研究可以扩展该方法到其他类似的灾难预测问题中，如航空事故生存预测或自然灾害中的人员安全预测。此外，结合其他机器学习和深度学习技术，如集成模型或迁移学习，可以进一步提高预测性能和泛化能力

## 6.2 研究贡献

本研究基于Keras深度学习框架，构建了一个用于预测泰坦尼克号灾难的生存模型。通过对数据的分析和模型训练，我们得到了准确的生存预测结果，为历史研究和船舶安全领域提供了有价值的参考。

我们也预见了深度学习在灾难预测中的潜力。本研究表明深度学习方法可以应用于灾难预测领域，为我们提供了一种新的角度来研究和分析历史事件中的生存情况。未来的研究可以将该方法扩展到其他类似的灾难预测问题中，以提高人员安全和预警系统的效能

## 6.3 后续工作

在进一步研究中，我们可以考虑扩大数据集，收集更多乘客和船舱的信息，以提高模型的泛化能力。此外，我们还可以探索其他的机器学习算法， 如随机森林， 梯度上升， 决策树等等与深度学习模型，以比较它们在灾难预测中的性能。

通过本研究，我们对泰坦尼克号灾难预测的问题进行了深入研究，并提出了一个基于Keras的生存预测模型。这个模型不仅可以帮助我们了解泰坦尼克号灾难中的生存机会因素，也为相关领域的研究和预防提供了有针对性的参考。

参考文献

[1]. Kaggle: Titanic: Machine Learning From Disaster. [Online]. Available: https://www.kaggle.com/c/titanic

[2]. Brownlee, J. (2020). How to Prepare Your Data for Machine Learning in Python. [Online]. Available: https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning-in-python/

[3]. Ekinci, E., Omurca, S. İ., & Acun, N. (2018, December). A comparative study on machine learning techniques using Titanic dataset. In *7th international conference on advanced technologies* (pp. 411-416).

[4]. Singh, A., Saraswat, S., & Faujdar, N. (2017, May). Analyzing Titanic disaster using machine learning algorithms. In *2017 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 406-411). IEEE.