一、研究背景与目的

1912 年 4 月 15 日，豪华邮轮泰坦尼克号在首航途中撞上冰山并沉没，造成 1500 余人遇难。作为历史上最著名的海难事件之一，其生还与死亡背后反映了复杂的社会结构与人类行为特征。乘客的舱位等级、性别、年龄及经济地位等因素可能在灾难中显著影响生存概率。本研究旨在利用决策树、朴素贝叶斯与逻辑回归三种算法，对泰坦尼克号乘客的生存概率进行建模分析，比较不同模型的性能表现，探索性别、舱位等级、票价与社会身份等因素对生存率的影响，从而为人类社会行为模式的定量研究提供经验与方法论参考。

二、数据处理与初步探索

2.1 数据结构与字段说明

本研究以 Titanic 数据集为研究对象，包含乘客的基础信息与生存标签。训练集规模为 891×12，测试集为 418×11，二者在特征结构上基本一致，仅测试集缺少目标变量 Survived。主要字段包括乘客编号（PassengerId）、生存状态（Survived）、舱位等级（Pclass）、姓名（Name）、性别（Sex）、年龄（Age）、船上兄弟姐妹或配偶数量（SibSp）、船上父母或子女数量（Parch）、船票编号（Ticket）、票价（Fare）、舱位号（Cabin）以及登船港口（Embarked）。其中，Survived 为目标变量，取值 1 表示幸存、0 表示未幸存；Pclass 表示乘客所处舱位等级，数值越小舱位等级越高；Sex 和 Age 分别反映性别与年龄，是影响生存率的重要特征；SibSp 与 Parch 则共同描述家庭关系与同行人数；Fare 为票价信息，能在一定程度上反映经济状况；Cabin 和 Embarked 提供了乘客的登船位置与舱位位置等空间信息。

总体而言，数据维度覆盖人口学特征、社会阶层以及登船位置，为后续生存预测模型提供了较为全面的输入基础。

训练集样本 891 条中，约 38.4% 乘客幸存，存在显著性别、生舱等级差异，表明这些特征可能与生存率存在强相关性。

2.2 缺失值与数据质量检查

在对数据进行缺失值统计与热力图可视化后发现，Cabin 列存在大比例缺失，约占 77%，推测原因在于仅部分舱位信息被记录。该特征的缺失可能导致噪声积累，因此后续仅保留其首字母作为简化变量 Deck，用于捕捉舱位层级特征。Age 列存在约 20%的缺失样本，而 Embarked 仅缺失 2 条记录（约 0.2%）。测试集中还发现 Fare 有 1 条样本缺失。总体来看，除 Cabin 外，数据的完整性较好，未出现系统性缺失。

三、数据预处理与特征工程

3.1 缺失值处理

①Embarked：作为分类变量，缺失比例极低，因此使用众数（S 港）进行填补；

②Fare：显著的右偏长尾分布，票价跨度大，尤其头等舱乘客票价远高于经济舱，个别样本超过 400 英镑，为避免均值受极端值影响，采用(Pclass, Embarked)组合分组的中位数进行填补，从而兼顾舱位与登船地点的经济层次差异；

③Age：呈轻度右偏分布，集中在 20 至 40 岁区间。根据箱线图结果，年龄的四分位距（IQR）约为 17 岁，上限边界约为 64 岁，下限接近 0 岁。经计算，共识别出 7 个高于上界的异常点，这些样本对应年长乘客（如 65 岁以上老年人），属于合理值而非录入错误，因此保留在数据集中采用(Title, Pclass, Sex) 组内均值填补；为兼顾数据连续性与模型假设，采用(Title, Pclass, Sex) 三维分组的均值进行填补。这种分组方式不仅考虑了社会身份与舱位差异，还能有效保留方差结构。

3.2 特征提取与构造

为了提升模型的表达能力和解释性，本研究进行了系统的特征工程：

①提取头衔：从姓名字段中提取称谓（如 Mr, Mrs, Miss, Master 等），并将稀有称谓合并为统一类别“Rare”；

②构造家庭特征：通过 SibSp+Parch+1 计算家庭成员总数 FamilySize，并由此衍生出 IsAlone；

③处理 Cabin 信息：提取 Cabin 首字母，缺失部分统一归类为“U”，以保留位置层次信息；

④处理 Ticket 信息：提取 Ticket 前缀为 TicketPrefix，同时统计相同票号乘客数量形成 TicketGroupSize，反映团体出行特征；

⑤平滑 Fare 票价：对票价进行对数变换，以降低偏态性并提升模型稳定性

3.3 编码

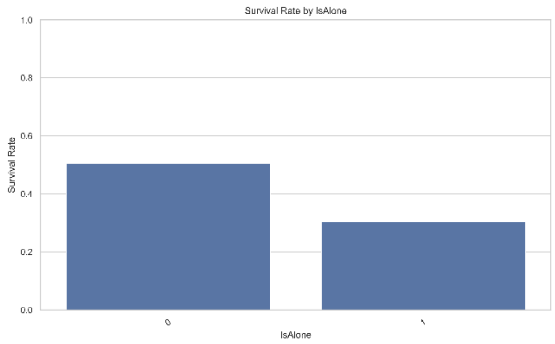
在特征工程完成后，所有分类变量均通过 One-Hot 编码转换为哑变量形式，确保模型能够处理非数值特征。例如 Sex 最初以文本形式记录（"male" 或 "female"），无法直接输入模型，故先将其转换为 Pandas 的分类变量，随后通过 One-Hot 编码生成两个互斥的二进制特征列，即 Sex_female 和 Sex_male。

四、探索性数据分析

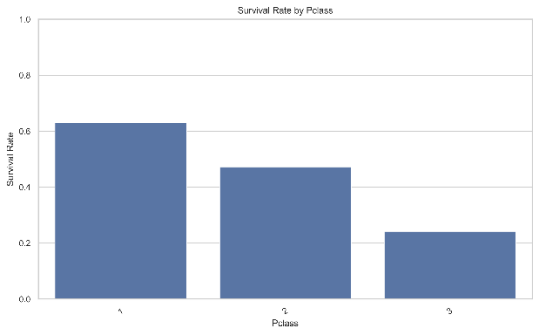
4.1 生存率与类别变量关系

为初步探究乘客属性与生存率之间的关系，本研究从性别、舱位等级、登船港口、称谓（Title）以及是否独自出行（IsAlone）等类别变量入手，绘制生存率柱状图并进行比较分析，结果如下：

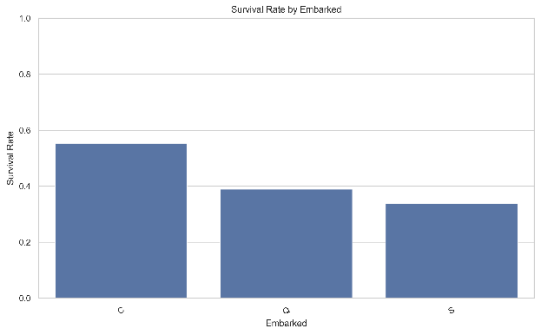
图表 1：乘客性别与生存率关系柱状图



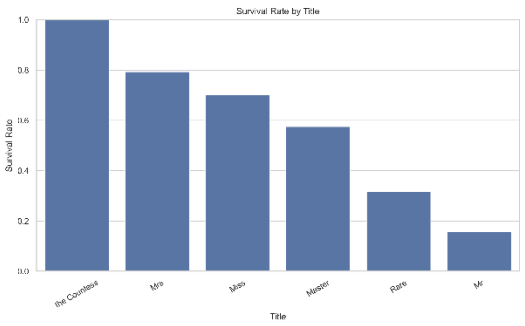
图表 2：各舱位等级乘客生存率比较



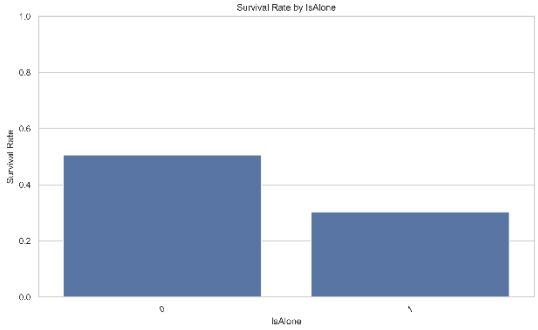
图表 3：不同登船港口乘客生存率对比



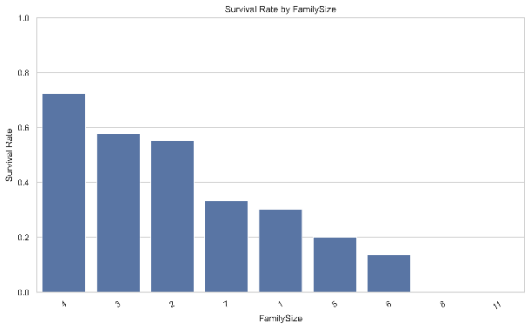
图表 4：称谓（Title）与乘客生存率关系图



图表 5：是否独自出行（IsAlone）对生存率的影响



图表 6：家庭大小对生存率的影响



①性别差异显著。女性乘客的平均生存率远高于男性，这一现象与“女士优

先”的救援原则相符，也说明性别是影响生存率的首要因素；

②舱位等级与生存率呈正相关。头等舱（一等舱）乘客的生还概率最高，而三等舱乘客生存率最低，这反映了不同社会经济层次在紧急救援中所受到的优先程度差异；

③登船港口差异亦有一定影响。从瑟堡（C 港）登船的乘客生存率略高于南安普敦（S 港）和皇后镇（Q 港），可能与不同港口上船乘客的舱位等级和票价分布有关；

④称谓特征反映社会身份差异。如 “Mrs” 和 “Miss” 组乘客的生存率高于 “Mr” 组，而称谓 “Master” （通常为儿童）生还率也显著偏高，进一步印证了救援中性别与年龄的双重影响；

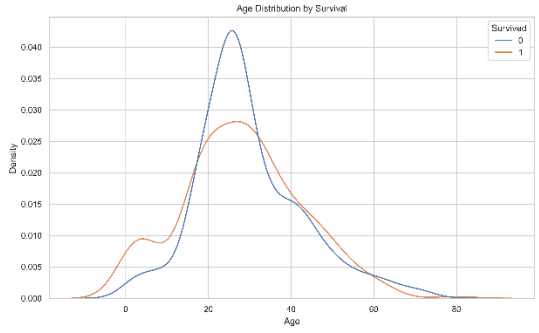
⑤家庭结构影响生存概率。家庭规模适中的乘客生存率最高，而独自旅行者和家庭成员众多者的生存率均相对较低，说明既有社会联系又不至于行动受限的乘客更容易获得救援。

这些分析表明，社会属性与群体身份变量对生存结果具有强烈的区分度，为后续建模提供了重要参考。

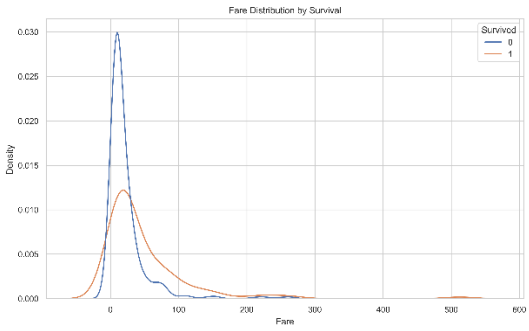
4.2 连续变量分布与生存情况

在连续型变量分析中，重点考察了乘客年龄（Age）与票价（Fare）的分布情况。通过 KDE 曲线对比生还与未生还两类样本，结果发现：

图表 7：乘客年龄分布及生存状态核密度曲线



图表 8：票价（Fare）分布及生存状态核密度曲线



①年龄分布方面，生还乘客的年龄整体偏低，集中在 20 岁以下和 30 岁左右区间，而未生还乘客的年龄分布更为分散。这表明年轻群体在救援过程中获得生存机会的可能性更高。

②票价分布方面，生还者的票价分布整体右移，说明票价较高（即舱位等级

较高）的乘客生存率更高。这与舱位等级特征的分析结果一致，也从数值层面验证了经济地位与生存概率之间的正向关系。

综上，连续变量的分布特征与类别变量分析结果相互印证，显示出明显的社会分层效应：高社会地位和较年轻的乘客在灾难中具有更高的生还概率。

4.3 特征相关性与自相关分析

为评估变量间的线性相关性，本研究计算了主要数值特征的 Pearson 相关系数矩阵。结果表明，多数特征之间的相关性较低，说明数据整体多维度、信息分布较为独立。

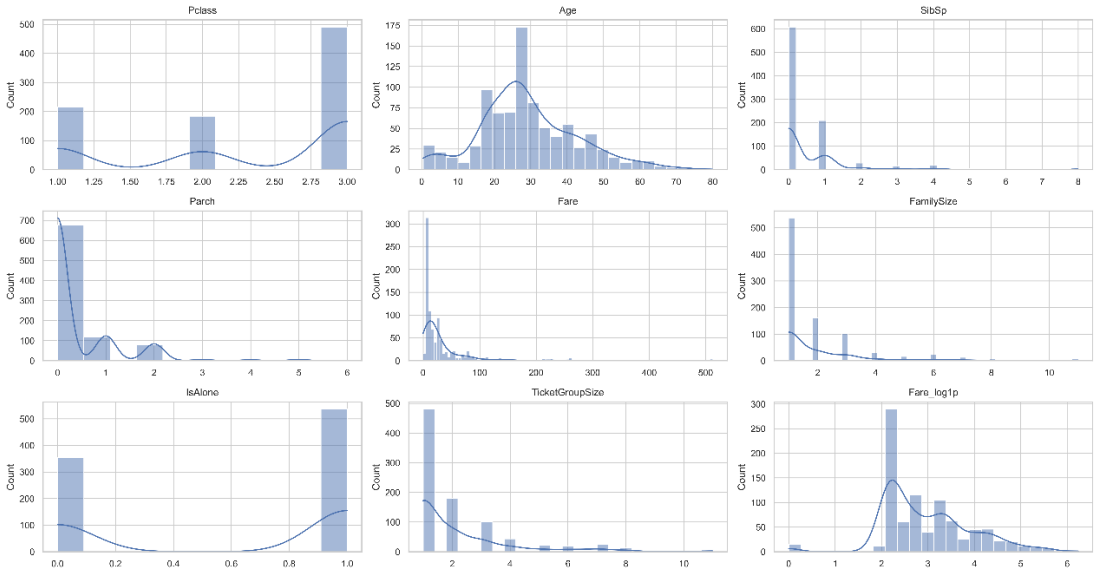
在核心特征中，Pclass 与 Fare 呈显著负相关（$r = -0.549$），反映舱位等级越高（数值越小），票价越高的事实；Age 与 Pclass 亦呈中度负相关（$r = -0.421$），表明头等舱乘客年龄略大。

FamilySize 与 SibSp（$r = 0.891$）及 Parch（$r = 0.783$）之间存在强正相关，而 IsAlone 与 FamilySize 呈高度负相关（$r = -0.691$），这一逻辑关系说明独自出行乘客的家庭规模必然较小。

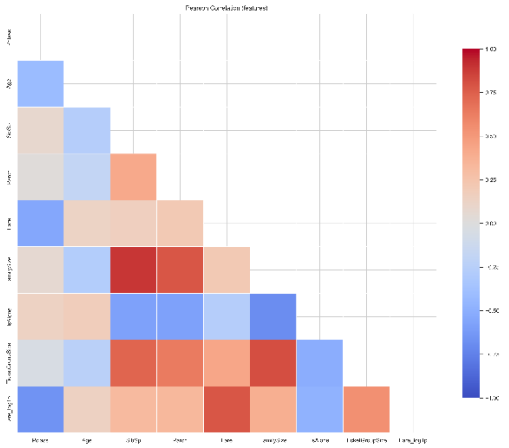
此外，Fare_log1p 与 Fare 的相关系数为 0.788，符合对数平滑后数值之间的函数依存关系。Sex_female 与其他数值特征相关性均较低（$|r|<0.3$），说明性别是独立于经济与家庭变量的重要社会特征。

总体来看，除家庭相关变量外，数据集中未出现严重的多重共线性，各主要变量之间的区分度良好。

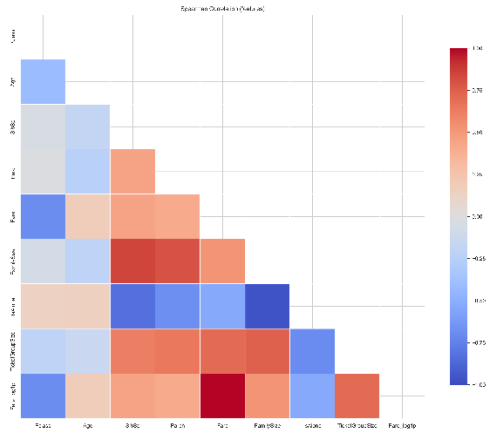
图表 9：自变量相关性检验

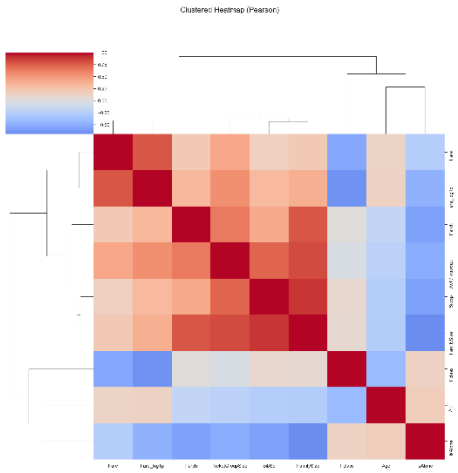


图表 10：主要数值变量 Pearson 相关热力图



图表 11：特征聚类热图（Spearman 相关）





五、模型构建与训练

5.1 数据集划分与评估标准

在建模阶段，采用 8:2 的比例将训练数据划分为训练集与验证集，并进行分层抽样（stratify=y），以保持生还与未生还样本比例一致。模型评估选取五个综合性指标：准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 值（F1 Score）以及受试者工作特征曲线下面积（ROC AUC）。其中，ROC AUC 能更全面地衡量模型在不同阈值下的分类能力，是本研究的主要比较指标。此外，训练过程中通过可视化呈现混淆矩阵、ROC 曲线、PR 曲线及学习曲线，直观评估各模型的预测性能与泛化能力。

5.2 决策树

在基线建模中，决策树模型采用 Gini 指数作为划分准则，并引入最大深度限制（max_depth = 4）以防止过拟合。模型在验证集上的表现为：

Accuracy=0.799， Precision=0.811，Recall=0.623，F1=0.705，AUC=0.831。此外，特征重要性分析显示，模型的决策主要依赖以下变量：

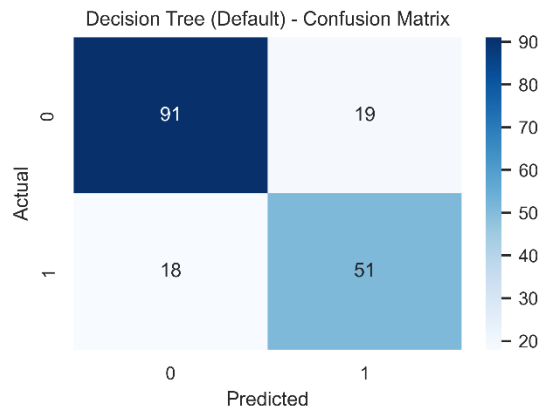
①Title_Mr 的重要性最高，为 0.540；

②其次为 Pclass（0.133 与 TicketGroupSize（0.097）；

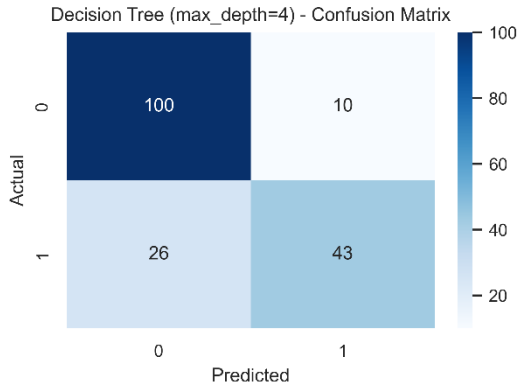
③其他影响较大的特征包括 Deck_U（0.078）、Title_Rare（0.053）以及 Age（0.041）。

这些结果表明，社会身份与舱位等级是决策树划分路径中的主导因素，票号分组特征则在一定程度上反映出团体乘客的生存优势。模型通过树形结构明确展示了不同条件下的生存概率分支，具备良好的可解释性。然而，由于决策树对局部样本划分较为敏感，其泛化性能略逊于参数化模型。

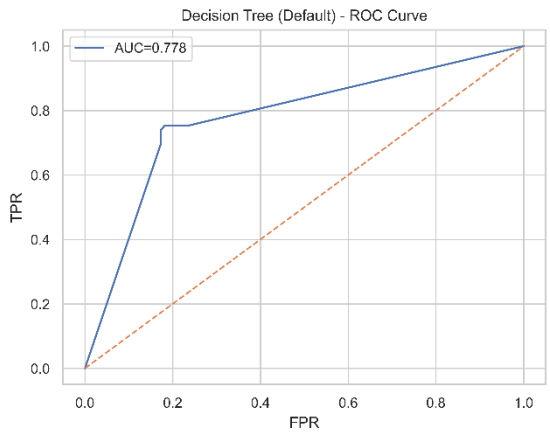
图表 12：决策树模型混淆矩阵



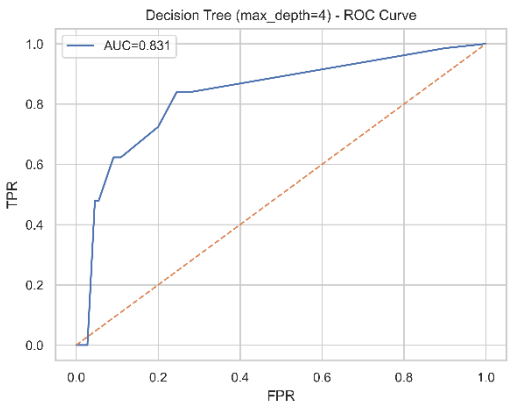
图表 13：决策树模型混淆矩阵（4）



图表 14：决策树模型 ROC 曲线

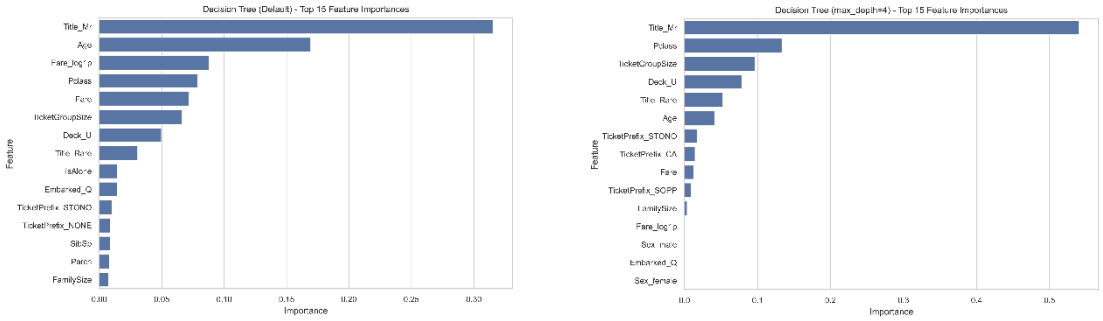


图表 15：决策树模型 ROC 曲线（4）

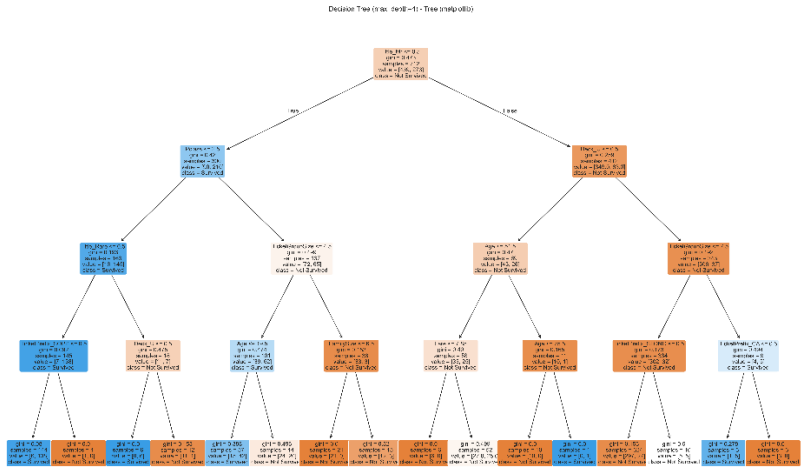


图表 16：决策树模型特征重要性条形图

图表 17：决策树模型特征重要性条形图（4）



图表 18：决策树结构

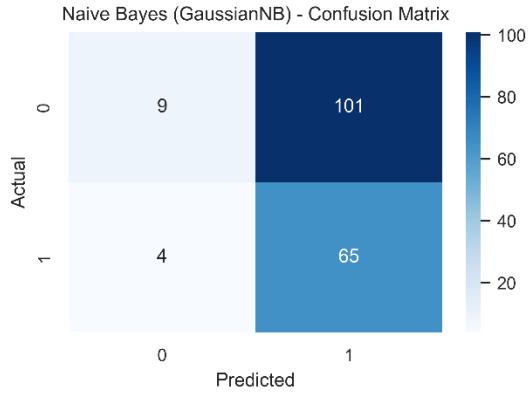


5.3 朴素贝叶斯

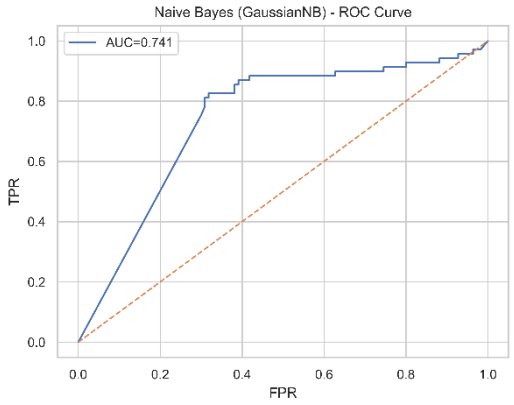
朴素贝叶斯模型基于特征条件独立的假设，并假定连续变量服从高斯分布。模型在验证集上的结果为：Accuracy=0.441，Precision=0.406，Recall=0.971，F1 Score=0.573，ROC AUC=0.814。可以看出，该模型召回率极高（0.97），即几乎能识别所有生还样本，但精确率偏低（0.41），说明预测中存在较多误报。这与模型独立性假设偏强有关：由于 Titanic 数据中多变量间存在一定相关性（如舱位与票价、家庭特征之间），朴素贝叶斯在联合分布建模时无法完全捕捉复杂关系，导致分类边界相对宽松。

尽管如此，该模型收敛迅速、计算效率高，适合作为基线对照模型或在高维稀疏特征情境下使用。

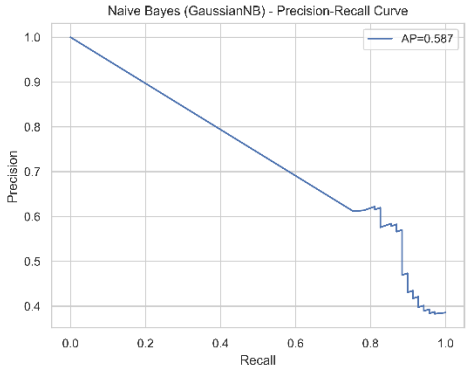
图表 19：朴素贝叶斯模型混淆矩阵



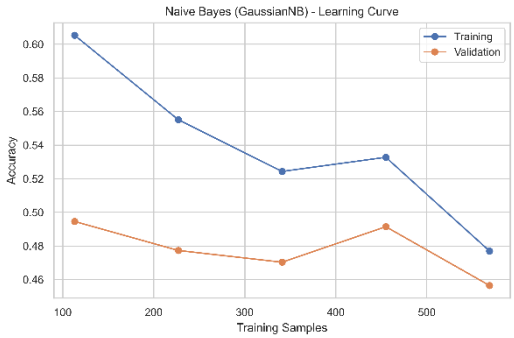
图表 20：朴素贝叶斯模型 ROC 曲线



图表 21：朴素贝叶斯模型召回曲线



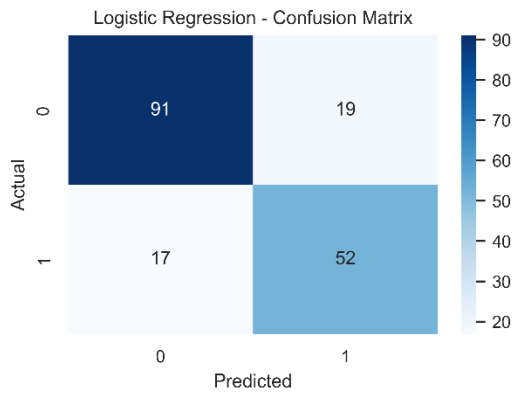
图表 22：朴素贝叶斯模型学习曲线



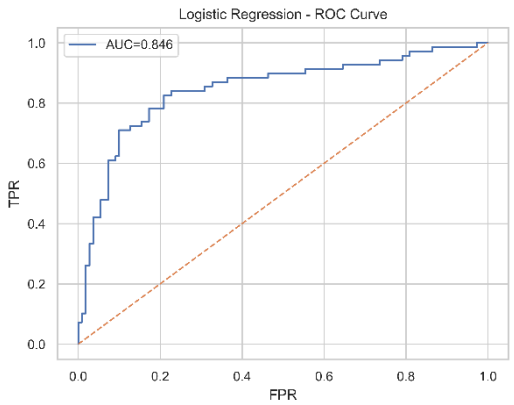
5.4 逻辑回归

逻辑回归模型采用标准化 Pipeline 结构，并设置最大迭代次数 500 以确保收敛稳定。模型在验证集上表现最优，各指标如下：Accuracy = 0.860，Precision = 0.823，Recall = 0.812，F1 Score = 0.818，ROC AUC = 0.908。

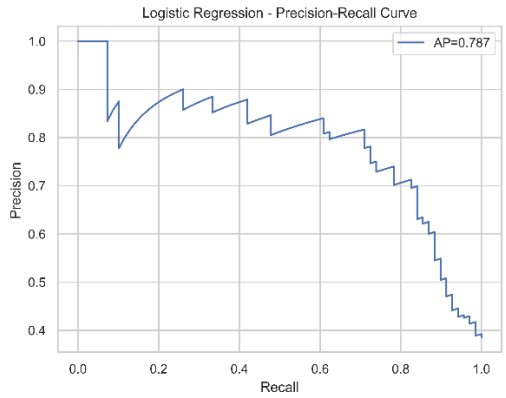
图表 23：逻辑回归模型混淆矩阵



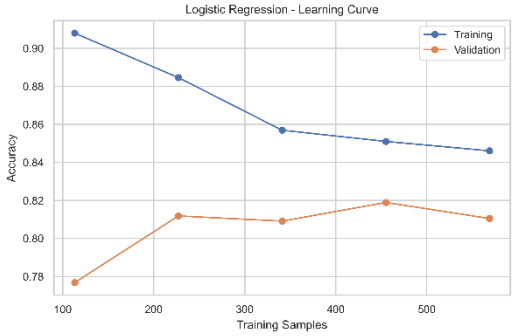
图表 24：逻辑回归模型 ROC 曲线



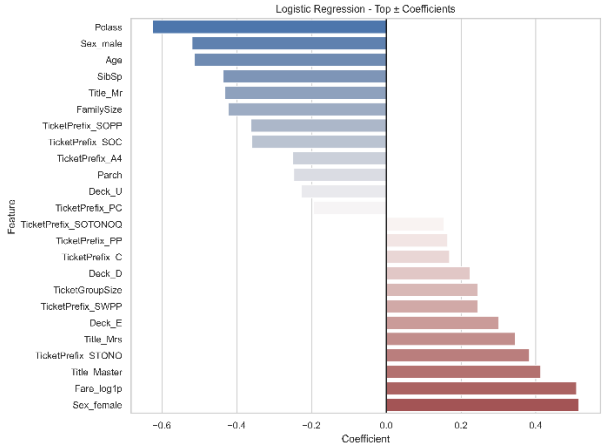
图表 25：逻辑回归模型召回曲线



图表 26：逻辑回归模型学习曲线



图表 27：逻辑回归模型前 10 个正负系数特征条形图



从系数分析结果看，影响生存概率的主要变量包括：

①正向影响因素（提高生还率）

Sex_female（0.558）、Fare_log1p（0.515）、Title_Master（0.515）、TicketPrefix_SWPP（0.353）、TicketPrefix_STONO（0.305）。

这些特征表明女性、儿童、票价较高及特定船票类别乘客的生还概率显著增加。

②负向影响因素（降低生还率）

Pclass（−0.673）、Sex_male（−0.561）、Age（−0.511）、SibSp（−0.484）、FamilySize（−0.432）

结果表明，舱位等级数值越高（即舱位越低）、男性乘客、年龄较大或家庭成员过多者，其生还概率均显著下降。

逻辑回归模型通过系数方向与大小清晰地揭示了各特征与生存率的线性关系，结果与探索性分析结论高度一致。该模型在稳定性和解释性上优于其他模型，ROC AUC 高达 0.908，为三者中表现最优。其良好的泛化性能与高区分能力表

明，该模型能较准确地刻画泰坦尼克号乘客的生存规律。

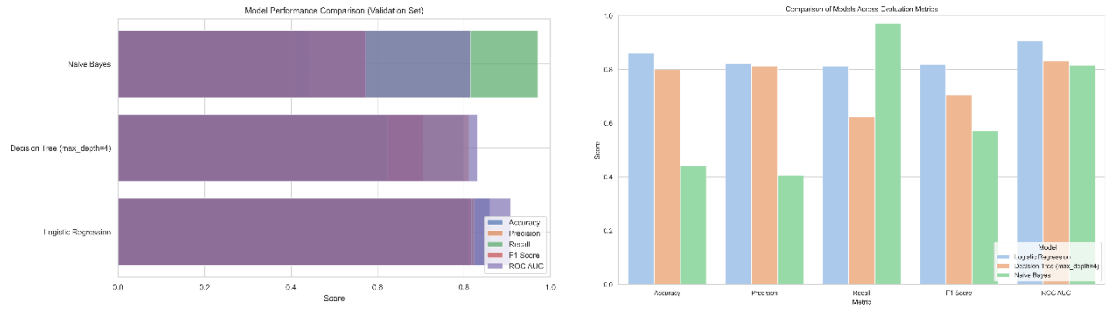
六、模型比较与结果总结

6.1 模型整体性能比较

本研究对三种分类模型（逻辑回归、决策树与朴素贝叶斯）进行了性能评估，验证集上的主要结果如下：

模型	准确度	精确率	召回率	F1 值	ROC AUC
逻辑回归	0.8603	0.8235	0.8116	0.8175	0.9080
决策树（MAX_DEPTH=4）	0.7989	0.8113	0.6232	0.7049	0.8312
朴素贝叶斯	0.4413	0.4061	0.9710	0.5726	0.8145

图表 28：三种模型主要指标对比图



逻辑回归模型的准确率最高，达到 0.860；决策树模型为 0.799，而朴素贝叶斯模型仅为 0.441。

在精确率方面，逻辑回归与决策树表现相近，分别为 0.823 与 0.811，说明两者在正例（生还者）的识别精度上均较高。

然而在召回率上，朴素贝叶斯表现出极高的灵敏度（0.971），但由于其精确率过低，造成大量误报，综合 F1 值仅为 0.573。相比之下，逻辑回归在召回率（0.812）与精确率之间取得平衡，其 F1 值最高（0.818），表明总体分类效果最佳。

从 ROC AUC 指标看，逻辑回归同样表现最优，达到 0.908，远高于决策树（0.831）和朴素贝叶斯（0.814）。这说明逻辑回归在不同分类阈值下均具有更高的区分能力和稳定性。

综合各项指标，逻辑回归在准确性、平衡性与判别能力方面均优于其他模型，是本研究的最优选择。

6.2 模型性能差异分析

从算法特性角度分析，三种模型性能差异主要由以下因素决定：

首先，逻辑回归属于线性模型，其优势在于能够通过参数估计精确刻画特征与生存概率之间的线性关系。由于泰坦尼克号数据集中性别、舱位等级和票价等变量与生存结果的关系近似线性，该模型能稳定收敛并输出具有可解释性的结果。其 ROC AUC 高达 0.908，说明模型在不同阈值下均能良好区分生还与未生还样本，且受噪声影响较小。

其次，决策树模型能捕捉非线性特征交互，对类别型变量较为敏感。其在本研究中的 AUC 为 0.831，说明模型能一定程度上刻画复杂的生存条件分支。但由于 Titanic 数据集样本量有限，若不加深度限制，树模型易出现过拟合。为抑制该问题，本研究将最大深度限制为 4，使模型在保持解释性的同时提升泛化性能。结果表明，决策树能清晰揭示性别、舱位等级和社会称谓等变量的关键作用。

最后，朴素贝叶斯模型在假设特征相互独立的条件下构建概率模型。该假设在 Titanic 数据中并不完全成立，例如舱位等级与票价高度相关（$r=-0.549$），家庭规模变量间亦存在强相关（$r>0.8$）。因此，朴素贝叶斯的过度独立性假设导致分类边界过于宽松，虽然召回率极高（0.971），能捕获几乎所有生还样本，但误报比例大幅上升，从而使准确率显著下降。这也解释了其 ROC AUC 虽达到 0.814，但整体可用性偏低。

6.3 特征作用与模型一致性验证

从特征权重与重要性角度来看，三种模型对主要变量的判断趋势保持一致：

①性别（Sex）：在所有模型中均为最显著的分类依据。逻辑回归中 Sex_female 系数为 +0.558，Sex_male 为 −0.561；决策树中该特征出现在顶层划分节点，显示性别对生存率的直接影响。

②舱位等级（Pclass）与票价（Fare）：均反映经济层次差异，相关系数 $r=-0.549$。两者在逻辑回归中系数符号相反，符合高等级舱位与高票价乘客生还率更高的现实规律。

③称谓（Title）与年龄（Age）：Title_Master 与低年龄群体在逻辑回归中表现为正向因素，说明儿童生还率较高。决策树的重要性分析亦将 Title_Mr（0.540）和 Age（0.041）列为关键分支变量。

家庭结构变量（FamilySize, IsAlone）：Pearson 相关系数显示二者高度负相关（$r=-0.691$）。在逻辑回归中，家庭规模系数为 −0.432，表明过大的家庭规模

可能在救援中造成不利影响，而完全独自旅行者同样风险较高。

6.4 预测结果与综合讨论

模型	预测幸存人数	幸存率
逻辑回归	178	42.6%
决策树 (MAX_DEPTH=4)	168	40.2%
朴素贝叶斯	152	36.4%

结果显示，三种模型对总体生还比例的预测均与历史真实比例（约 38.4%）接近，其中逻辑回归预测的幸存人数略高但仍处于合理范围内，反映出模型未出现明显的过拟合或系统偏差。而将三模型的预测结果进行一致性检验后则发现，约 82% 的乘客样本在三种模型中预测结果一致，说明模型在主要特征识别上具有较高稳定性。差异主要集中在票价处于中间区间（20–50 英镑）和家庭成员较多的样本中，这类乘客往往位于决策边界附近。

从整体表现来看，逻辑回归在准确率、AUC 和 F1 值上均优于其他模型，表现最为稳定。决策树在可解释性方面具有优势，可直观揭示特征分层结构，适合作为辅助分析模型。朴素贝叶斯则凭借高召回率适合在“尽量不漏检”的场景下使用，但其精确率较低，难以作为最终分类器。

综上，本研究选择逻辑回归模型作为最终的乘客生存预测模型。该模型不仅在验证集上获得最高的 ROC AUC（0.908），而且具备清晰的统计解释性与较强的泛化能力，能够较为准确地反映泰坦尼克号灾难中社会性别、经济地位与生存概率之间的系统性关系。

七、改进方向

本研究基于逻辑回归、决策树与朴素贝叶斯三种模型完成了对泰坦尼克号乘客生存预测的建模分析。虽然逻辑回归模型在验证集上取得了最优表现（Accuracy＝0.860，AUC＝0.908），但在特征丰富度、模型复杂度和评估方法上仍存在进一步提升空间。未来的研究可从以下几个方面加以改进与拓展。

首先，在特征拓展方面，可结合原始乘客名册信息，引入家庭 ID（FamilyID）以标识同户乘客，进一步刻画群体行为特征；同时对舱室（Cabin）位置进行分区编码，区分靠近甲板、船尾或船首的乘客，以反映不同逃生路径的潜在影响。此外，登船时间（EmbarkTime）和登船港口组合变量也值得纳入，以更精确地捕捉乘客分布与登船顺序间的社会层级差异。

其次，在模型优化方面，可在逻辑回归中引入 L1/L2 正则化，以降低高相

关特征对系数估计的干扰；同时尝试集成学习算法，如随机森林（Random Forest）、梯度提升树（XGBoost）或 LightGBM，以建模非线性特征交互。相较于单一模型，这些方法能在保持可解释性的同时，提升预测稳定性与鲁棒性。

再次，在模型融合层面，可采用多数投票或加权平均策略，将逻辑回归的线性可解释性与树模型的非线性判别能力相结合，从而提升整体泛化性能与决策可靠性。

最后，在评估优化方面，未来可采用 K 折交叉验证或分层采样验证方法，以减少数据划分带来的偶然性误差；同时引入精确率-召回率曲线与 KS 值等多维指标，实现对模型性能更全面的评估。

通过上述改进方向的实施，模型不仅能在预测精度上进一步提升，也能更深入地揭示灾难情境下社会结构、性别与经济因素之间的复杂互动机制。