基于机器学习的电影票房预测分析

**摘要**

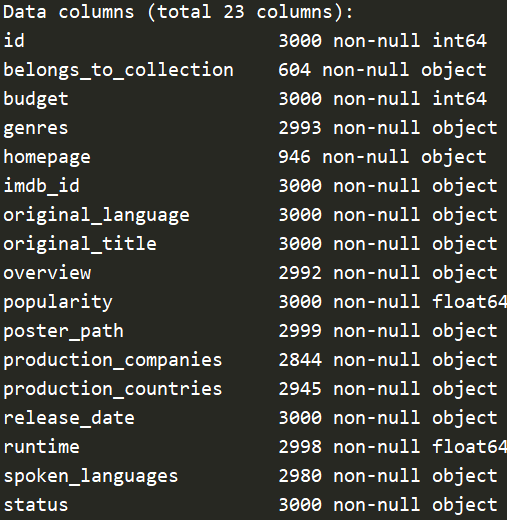
…….

**介绍**

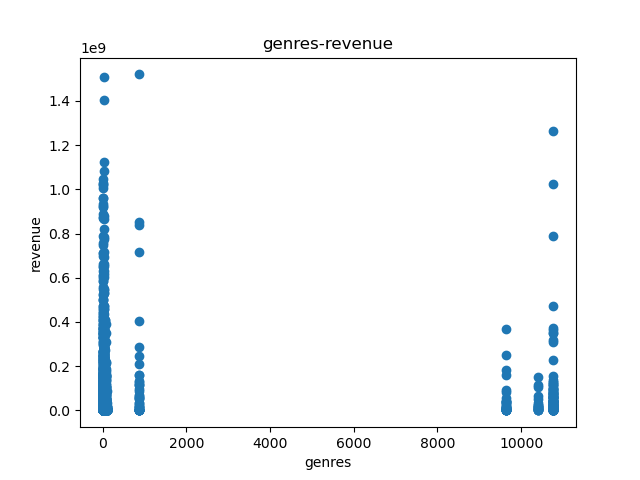
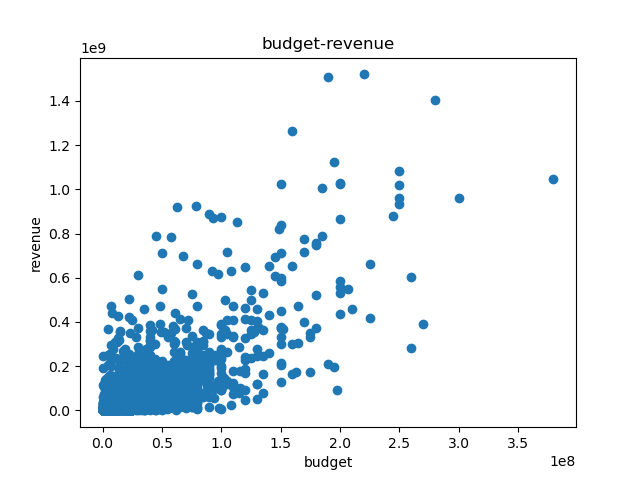
近几十年来,电影行业飞速发展,不论是在影片数量还是制作成本上,均呈现逐年递增趋势。宏大的电影市场规模、高昂的电影制作成本,使得票房预测成为电影行业迫切关注的问题。它是考察影响电影票房的诸多因素基础上，采用回归统计分析等方法研建立数学模型并预测。本文研究内容主要基于机器学习技术设计几种常见的电影票房预测模型并分析比较他们的优点，主要贡献包括三部分：1，电影票房的数据收集，包括类型，任务，预算等；2，电影特征抽取方法；3，基于机器学习的电影票房预测，预测未上映的电影中国票房收入，并分析比较这些方法的异同。

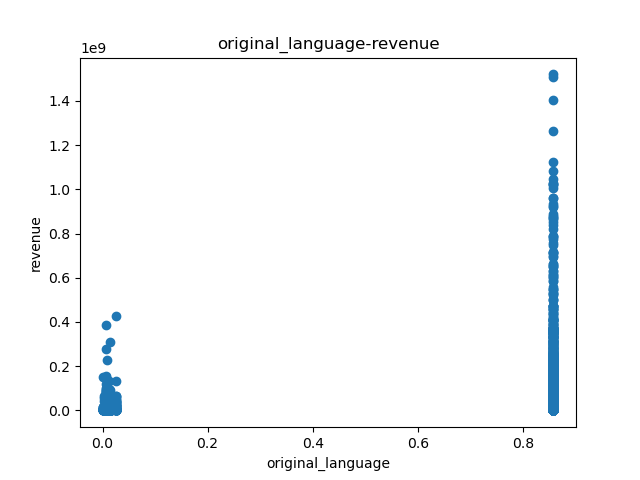
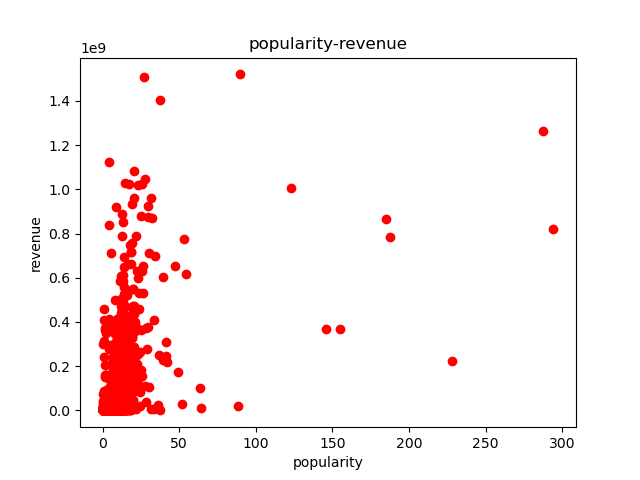
**数据收集和清洗**

我们的数据来源imdb电影网址，共有3000条数据，共有23列特征，具体特征见下图：

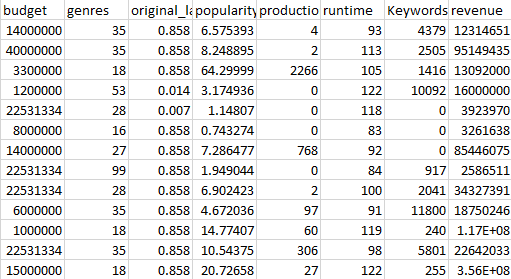


我们发现其中7个特征与票房相关性比较大，下面其中部分展示，票房和预算呈较明显的正相关关系。这很符合常识，





对于预算这一个特征，我们将其中数据为空的填充为整列数据的平均值，以便于后面的计算，对于电影类型和关键词特征，原始数据集的数据是字符串，而且有的字段还带有多个类型或者关键词，我们选择第一个电影类型或者关键词作为它的特征。而语言的选择也是票房的关键因素，我们从上图可以发现当电影语言为英语的时候票房比较高，所以我们将该票房的语言在整列语言中的所占的比例作为其特征值，这样可以精确的反应语言的欢迎程度。其他特征处理同上，经过以上处理，我们可以得到全为数值的特征：



接下来就是数据的划分，我们将数据按照7：3比例分为训练集和测试集。

**模型**

**多元线性回归：**

回归分析是机器学习一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。这种技术通常用于发现两个或多个变量之间的因果关系：1，表明自变量（预算/评分等）和因变（票房收入）之间的显著关系；2，表明多个自变量对一个因变量的影响强度。

1.把所有特征用pandas读取数据存储的.csv，写一个函数把数据转换为Train\_X, Train\_y, Test\_X, Test\_y.

2，用sklearn构造线性回归分析模型，其中predict为要预测的票房，函数返回对应的票房收入并计算与Test\_y的均方差，并返回模型预测时间。

**决策回归树：**

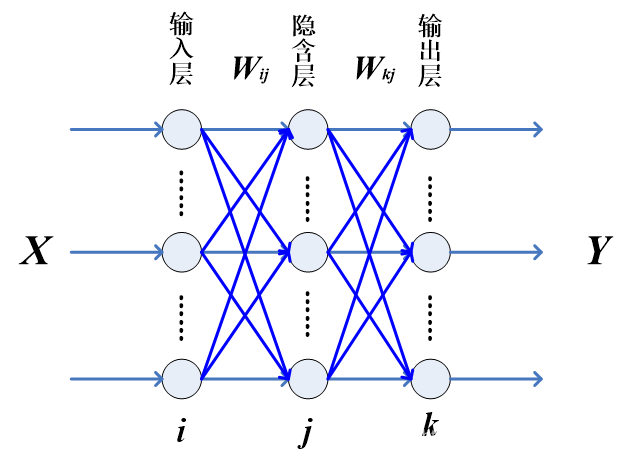
决策树模型是运用于分类以及回归的一种树结构。决策树由节点和有向边组成，一般一棵决策树包含一个根节点、若干内部节点和若干叶节点。决策树的决策过程需要从决策树的根节点开始，待测数据与决策树中的特征节点进行比较，并按照比较结果选择选择下一比较分支，直到叶子节点作为最终的决策结果。

1.把所有特征和目标用pandas读取数据存储的.csv，写一个函数把数据分为为Train\_X, Train\_y, Test\_X, Test\_y.

2，用sklearn构造决策回归分析模型，预测Test\_x，得到预测标签，函数返回对应的票房收入并计算与Test\_y的均方差，并返回模型预测时间。

**神经网络回归：**

最简单神经网络模型包括其输入输出模型、作用函数模型、误差计算模型和自学习模型。



输入层：输入神经元定义数据挖掘模型所有的输入属性值以及概率。一个感知器可以接收多个输入（x\_1,?\_2 ……?\_?），每个输入上有一个权值，此外还有一个偏置项b，就是上图中的?\_0。

隐含层：隐藏神经元接受来自输入神经元的输入，并向输出神经元提供输出。隐藏层是向各种输入概率分配权重的位置。

输出层：输出神经元代表数据挖掘模型的可预测属性值

激活函数：所谓激活函数（Activation Function），就是在人工神经网络的神经元上运行的函数，负责将神经元的输入映射到输出端,如sigmod函数、tanh函数等。

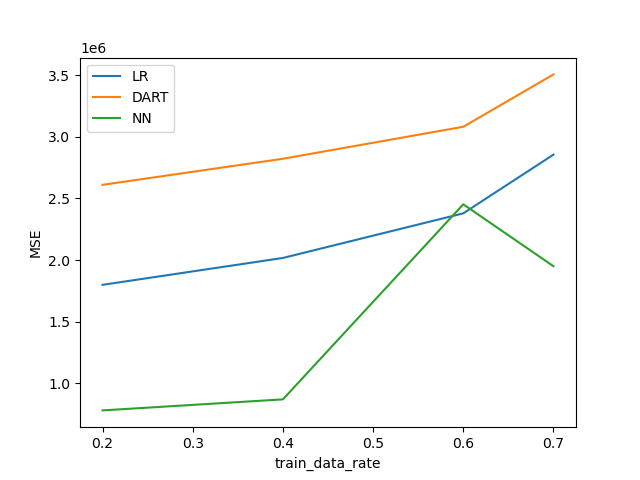
本文利用多层感知机对电影数据进行训练，下表是训练参数：

|  |  |
| --- | --- |
| 参数 | 值 |
| activation | tanh |
| hidden\_layer\_sizes | 2 |
| Batch\_size | 200 |
| Learning rate | 0.0001 |

对于隐藏层的层数，我们通过网格收索的方式寻找最好表现的值。训练时间比较长，返回测试结果并计算均方差和预测时间。

**实验结果和讨论分析**

|  |  |  |
| --- | --- | --- |
| Method | MSE | time |
| LR | 2855316 | 0.004 |
| DART | 3505899 | 0.0249 |
| NN | 1950035 | 41.4388 |

从上面结果我们可以发现神经网络在做回归分析是，由于其网络的复杂性，预测时间相对于其他两种方法非常长，但是预测结果却异常的好，这归功于神经网络结构的复杂性，对于一些非线性的数据能够很好的拟合,而决策回归树的结果表现最差，可能原因是该电影数据都是一些离散的，而对于决策回归树来说，它的强项是做分类任务，而线性回归技术通常用于发现两个或多个变量之间的因果关系，当自变量与因变量直接的关系比较简单时，该模型表现比较好，但是特征较多，变量之间存在非线性关系，该模型可能表现不好。

当我们逐渐改变训练集和测试集的比例，也就是不断增大训练集，我们可以发现前两个算法趋于稳定，mse变化不是很大，而第三个神经网络算法变化幅度比较大，不稳定，随着训练数据的增多，误差变得越来越多，模型精度也逐渐减小，但是总体来说神经网络的精确度明显高于其他两个算法。

**总结与展望**

电影票房预测在实际中具有实际意义，它能分析预测不同种类电影的票房价值，成为电影产业投融资重要参考工具，对电影产品定价及衍生产品开发都具有较强的指导作用。抽取特征后可能会出现聚类不明显的情况，还需对数据进行降噪、过滤处理，增加神经网络训练次数。电影市场潜力巨大，随着模型的不断完善，也会应用到其他领域。