**企业财务数据分析与造假识别**

**摘要**

财务数据是指企业经营活动和财务结果的数据记录，反映了企业的财务状况与经营成果。对行业、企业进行财务数据分析，就是要评价其过去的经营业绩、衡量现在的财务状况、预测未来的发展趋势。财务数据分析不仅对企业生产经营管理有着重要的作用，而且对企业外部投资、贷款、赊销等决策有重要作用；经济管理部门也可以通过财务数据分析了解行业经济的运行状况，合理配置资源，科学调控经济运行，促进经济高质量发展。而财务造假则会对企业、投资者和整个市场带来极大的危害，准确甄别财务造假企业，帮助投资者识别风险，促进企业诚信经营，对营造良性的市场竞争环境有非常重要的现实意义。总之，准确的财务数据分析对于企业和市场的稳健发展至关重要。而通过及时发现和应对潜在的财务问题，可以有效维护市场的公正竞争环境。

**针对任务一：**主要是对财务数据进行有效的数据预处理，以建立清洗完善的数据集。通过逐步的数据读取、合并、清理和格式转换，得到具备可分析性的“”文件。该文件不仅包含了经过有效处理的数据，还添加了利润率和资产负债率等指标，为后续分析提供了有力支持。得到的相关结果见附件“”、“”、“”、“”、“”、

“”。

**针对任务二：**主要是对财务数据的深入分析和可视化展示。通过对数据的统计和图表绘制，详细呈现了行业营业利润对比分析和行业企业营收分析结果。最终整合为一个大屏“行业与企业营业数据分析”，更加直观地展示了财务数据的关键指标，为深入理解行业现状和企业表现提供了全面视角。得到的相关可视化结果见正文。

**针对任务三：**主要是对企业利润预测和财务造假的识别。通过建立预测模型，成功预测了企业的利润总额，为企业未来经营提供了参考。同时，通过对财务数据的分析和识别特征的挖掘，成功筛查出了潜在的财务造假企业，为投资者提供了风险识别的重要线索。整体而言，任务三通过深入挖掘数据的关联性和特征，为企业提供了有力的财务决策支持。得到的相关结果见正文。

**Abstract**

Financial data refers to the recorded information of a company's operational activities and financial results, reflecting the financial condition and operating outcomes of the enterprise. Conducting financial data analysis for industries and companies is aimed at evaluating past operational performance, measuring current financial conditions, and predicting future development trends. Financial data analysis plays a crucial role not only in the operational management of companies but also in external decisions such as investments, loans, and credit sales. Additionally, economic management departments can utilize financial data analysis to comprehend the operational status of industries, allocate resources reasonably, regulate economic operations scientifically, and promote high-quality economic development. Financial fraud poses significant risks to companies, investors, and the entire market. Accurate identification of companies engaged in financial fraud is crucial for helping investors recognize risks and promoting honest business practices, contributing significantly to creating a positive market competition environment. In summary, accurate financial data analysis is vital for the robust development of both companies and markets. Timely detection and response to potential financial issues effectively maintain a fair competitive market environment.

**For task one:** Task One primarily involves effective preprocessing of financial data to establish a cleaned and comprehensive dataset. Through sequential data reading, merging, cleaning, and format conversion, the analyzable file "LR\_new.csv" is obtained. This file not only includes data that has been effectively processed but also incorporates key indicators such as profit margin and debt-to-assets ratio, providing robust support for subsequent analysis. For the relevant results, refer to the attached files“LR\_1.csv”、“LR\_2.csv”、“LR\_3.csv”、“LR\_4.csv”、“LR\_5.csv”、

“LR\_new.csv”.

**For task two:** Task Two focuses on in-depth analysis and visualization of financial data. Through statistical analysis and charting, the detailed results of industry operating profit comparison and industry enterprise revenue analysis are presented. These results are integrated into a large screen named "Industry and Enterprise Operating Data Analysis," providing a more intuitive display of key financial indicators for a comprehensive understanding of industry status and enterprise performance.Refer to the following sections for the related visualizations.

**For task three:** Task Three primarily involves enterprise profit prediction and the identification of financial fraud. By establishing a predictive model, the total profit for enterprises is successfully predicted, providing valuable references for future business operations. Simultaneously, through the analysis of financial data and the exploration of fraud identification features, potential cases of financial fraud are successfully identified, offering crucial clues for investors to recognize risks. Overall, Task Three provides robust financial decision support for companies through in-depth exploration of data correlation and features. Refer to the main body of the text for the detailed results.

**第1章 问题分析**

**1.1问题重述**

（1）对财务数据进行预处理。读取所给数据，对表中数据进行筛选、合并、删除空值、转换格式、插入数据等操作并计算相应的指标。

（2）分析财务数据指标并进行可视化。统计表中数据，进行行业营业利润对比分析和行业与企业营收分析，并绘制柱状图，折线图，饼图和柱状折线合图展现。

（3）建立企业利润预测模型，预测企业营收情况；并根据财务报表数据的特点，筛查财务造假的企业。计算指标间的相关性，进行指标的挑选以建立企业利润预测模型。利用关键因子分析样本数据，挖掘财务造假的识别特征。

**第2章 财务数据处理**

**2.1任务1.1-1.6数据预处理**

本问需要通过读取所给数据，对表中数据进行筛选、合并、删除空值、转换格式、插入数据等操作并计算相应的指标，以建立清洗完善的数据集，方便后续的数据分析和模型建立。

**2.1.1任务1.1：筛选字段**

本问首先需要引入Python中的pandas库，读取“”的数据，提取数据指标，筛选出字段“” 值为“A”的数据，将筛选出的数据另存为文件“”（文件编码设置为UTF-8），筛选后的数据行数为33415行 、列数为63列，呈现部分如下表1所示：

**表1 字段筛选**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | …… | B006000101 | B006000102 |
| 600696 | 2018/3/31 | A | …… | 6730614.54 | 330203.02 |
| 547 | 2018/3/31 | A | …… | 78689795.49 | 5138570.87 |
| 2772 | 2018/3/31 | A | …… | 88248093.13 | 1285531.73 |
| …… | …… | …… | …… | …… | …… |
| 600759 | 2019/1/1 | A | …… | 567712492.7 | -8379376.86 |
| 570 | 2018/12/31 | A | …… | 163314868.4 | 140548473.4 |
| 200570 | 2018/12/31 | A | …… | -188253346.5 | 145323.8 |

**2.1.2任务1.2：合并数据**

本问要读取“”、“”、“”三个数据文件。根据“”、“”和“”三个字段，提取“”中字段为“A002000000”和“A001000000”的相应数据，合并到“”中。根据字段“”，提取“”中字段为“”和“”的相应数据，合并到“”中。将完成合并的数据另存为文件“”（文件编码设置为 UTF-8）。重要代码如下：

li1,li2,li3,li4 = [],[],[],[]  
for i in range(len(data1)):  
 v1 = data2[(data2["Stkcd"] == data1.loc[i,"Stkcd"]) & (data2["Accper"] == data1.loc[i,"Accper"]) & (data2["Typrep"] == data1.loc[i,"Typrep"])]  
 li1.append(v1["A002000000"].values[0])  
 li2.append(v1["A001000000"].values[0])  
 v2 = data3[(data3["Stkcd"] == data1.loc[i,"Stkcd"])]  
 li3.append(v2["Indnme"].values[0])  
 li4.append(v2["Nindnme"].values[0])  
data1["A002000000"] = li1  
data1["A001000000"] = li2  
data1["Indnme"] = li3  
data1["Nindnme"] = li4

合并后数据的行数为33415行、列数为67列，呈现部分如下表2所示：

**表2 数据合并**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  | …… |  |  |
| 600696 | 2018/3/31 | A | …… | 房地产 | 房地产业 |
| 547 | 2018/3/31 | A | …… | 工业 | 信息技术业 |
| 2772 | 2018/3/31 | A | …… | 综合 | 农业 |
| …… | …… | …… | …… | …… | …… |
| 600759 | 2019/1/1 | A | …… | 工业 | 石油和天然气开采业 |
| 570 | 2018/12/31 | A | …… | 工业 | 普通机械制造业 |
| 200570 | 2018/12/31 | A | …… | 工业 | 普通机械制造业 |

**2.1.3任务1.3：删除空值占比高的列**

本问要读取“”，删除空值占比达70%及以上的数据列，将处理后的数据另存为文件“”（文件编码设置为 UTF-8），处理后数据的列数为34列，重要代码如下：

n\_num = data4.isnull().sum(axis=0).to\_list()  
yu\_num = len(data4)\*0.7  
name = data4.columns.to\_list()  
li1 = []  
for i in range(len(n\_num)):  
 if n\_num[i] >= yu\_num:  
 li1.append(name[i])  
df2 = data4.drop(li1,axis=1)

**2.1.4任务1.4：删除含空值的行**

本问要读取“”，删除包含空值的行，将处理后的数据另存为文件“”（文件编码设置为 UTF-8），处理后数据的行数为4742行。代码如下：

df3 = data5.dropna(axis=0, how='any')

**2.1.5任务1.5：转换数据类型**

本问要读取“”，将字段“”的日期数据转换为“”的格式，将处理后的数据另存为文件“”（文件编码设置为 UTF-8）。重要代码如下：

lr\_4 = pd.read\_csv(r'./数据/LR\_4.csv')

lr\_4['Accper'] = pd.to\_datetime(lr\_4['Accper'], format='%Y-%m-%d')

lr\_4.to\_csv(r'./数据/LR\_5.csv',encoding='utf-8')

处理后的数据呈现部分，如下表3所示：

**表3 日期数据转换**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| 600705 | 2018-03-31 | A |
| 921 | 2018-03-31 | A |
| 2286 | 2018-03-31 | A |
| 546 | 2018-03-31 | A |
| 600710 | 2018-03-31 | A |
| 2168 | 2018-03-31 | A |

**2.1.6任务1.6：计算指标并插入处理数据**

本问要读取“”，插入“利润率”和“资产负债率”两列。根据下表公式，计算对应的利润率和资产负债率，追加到“”对应字段。重要代码如下：

ind\_li = []  
lr\_li = data6["B001000000"]/data6["B001100000"]  
zc\_li = data6["A002000000"]/data6["A001000000"]

分别删除表中利润率、资产负债率不在[-300%，300%]范围内的行，将处理后的数据另存为文件“”（文件编码设置为 UTF-8）。重要代码如下：

for i in range(len(zc\_li)):  
 if zc\_li[i] < -3 or zc\_li[i] > 3 or lr\_li[i] < -3 or lr\_li[i] > 3:  
 ind\_li.append(i)  
data6["利润率"] = lr\_li  
data6["资产负债率"] = zc\_li  
df3 = data6.drop(ind\_li)

处理后的数据行数为4736行、列数36列，其中，前 5 个企业的利润率和资产负债率如下表4所示：

**表4 前 5 个企业的利润率和资产负债率**

|  |  |  |
| --- | --- | --- |
|  | 利润率 | 资产负债率 |
| 其他金融业 | 0.488310261 | 0.871995361 |
| 电器机械及器材制造业 | 0.040118928 | 0.674484109 |
| 食品加工业 | 0.043277884 | 0.329235179 |
| 其他社会服务业 | 0.019305578 | 0.507984412 |
| 批发和零售贸易 | 0.016178514 | 0.834009011 |

**第3章 财务数据指标分析及可视化**

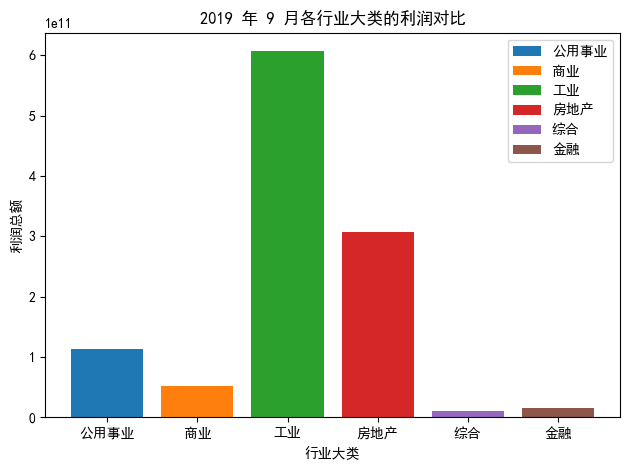
**3.1任务2.1-2.3可视化分析**

本问通过读取处理完成的表，进行数据统计，绘制“行业营业利润对比分析”图和2019年9月营业利润率均值排名第1的行业大类相关的“行业与企业营收分析”图，最后将绘制的图整合成“行业与企业营业数据分析”，并以可视化的形式呈现。

**3.1.1任务2.1：行业营业利润对比分析**

**（1）绘制各行业利润对比柱状图**

通过读取“”，分别统计不同行业大类 2019 年 9 月利润总额的均值，得到的2019年9月各行业大类的利润对比图如下图1所示：

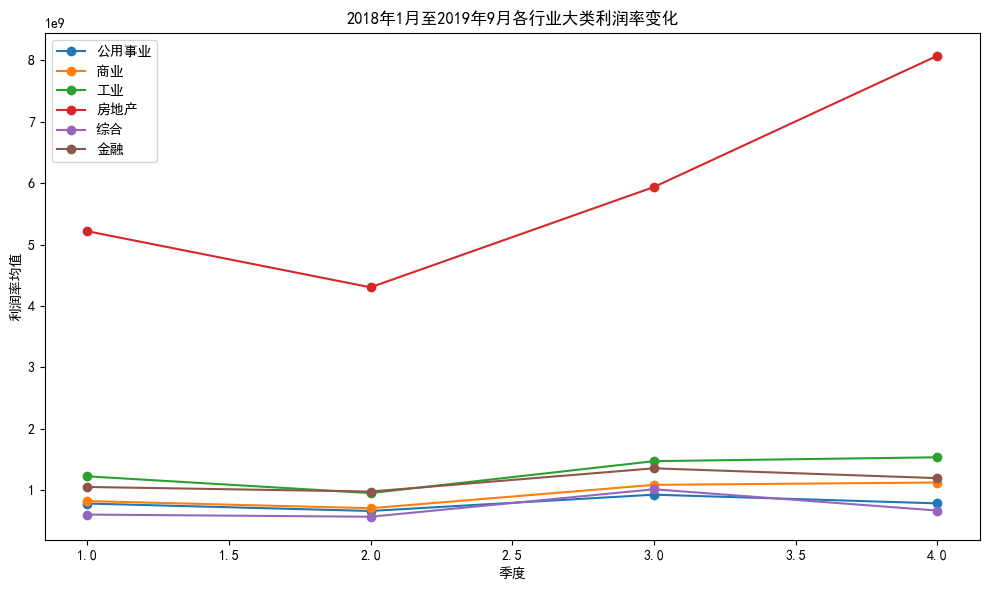
****

**图1 2019年9月各行业大类的利润对比**

在图1中，通过对比2019年9月不同行业大类的利润总额的均值，我们发现各行业利润差距十分显著，尤其是工业利润最高，其次是房地产行业，其他行业利润总额相对较低。

**（2）绘制各行业利润率变化折线图**

通过统计不同行业大类2018年1月至2019年9月各季度利润率均值，得到各行业大类利润率均值变化折线图如下图2所示：

****

**图2 2018年1月至2019年9月各行业大类利润率变化**

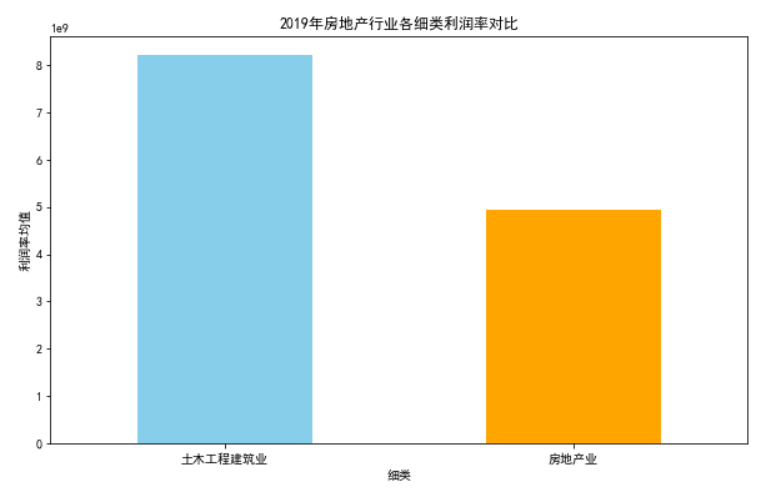
图2清晰的展现了2018年1月至2019年9月各行业各季度的利润率均值的变化趋势，公用事业、商业、工业、综合类行业、金融业变化趋势大致相同，都是在第三季度左右达到利润率均值的峰值，但其中工业的利润率一直不低于其他四类行业。相对于房地产行业来说，这些行业的利润率变化趋势较为稳定，房地产行业利润率波动性较大，第一季度到第二季度呈下降趋势，第二季度到第四季度一直处于上升趋势，但房地产行业的利润率一直远远高于其他所有行业。

**3.1.2任务2.2：行业与企业营收分析**

本问通过读取“”，可确定2019年9月营业利润率均值排名第1的行业大类为房地产。

**（1）绘制房地产行业各细类利润率对比柱状图**

通过统计房地产行业不同细类2019年9月的利润率，绘制出的利润率排名前3细类的利润率柱状图如下图3所示：

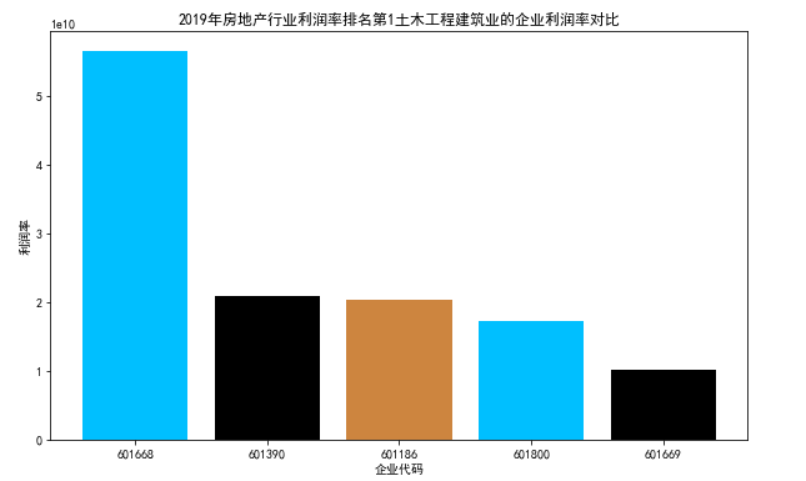


**图3 2019年房地产行业各细类利润率对比**

通过对比2019年房地产各细类行业的利润率均值可得，土木工程建筑业、房地产业利润率较高，其中土木工程建筑业利润率最高并远高于第2名，处于遥遥领先的地位。

**（2）绘制房地产行业利润率第一的细类的企业利润率对比柱状图**

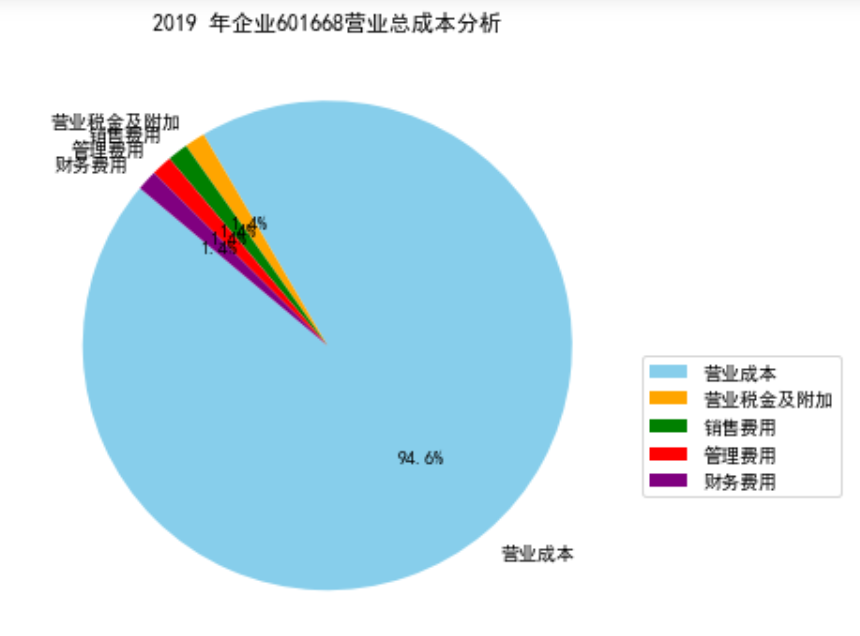
由图3可知房地产行业大类中利润率排名第1细类为土木工程建筑业，对土木工程建筑业各企业2019年9月利润率进行排序，绘制出排名前5企业的利润率柱状图如下图4所示：

 **图4 2019年土木工程建筑业的企业利润率对比**

其中601668产业利润率最高，601390、601186以及601800利润率其次，最差的是601669。由于题中要求将利润率排名第1的企业记为“T1”，故“T1”是企业代码为601668的企业。

1. **绘制企业“T1”营业总成本饼图**

由图4可知“T1”是企业代码为601668的企业，通过统计该企业2019 年 9 月财务报表的营业成本、营业税金及附加、销售费用、管理费用、财务费用，绘制出该企业2019年营业总分析饼图如下图5所示：

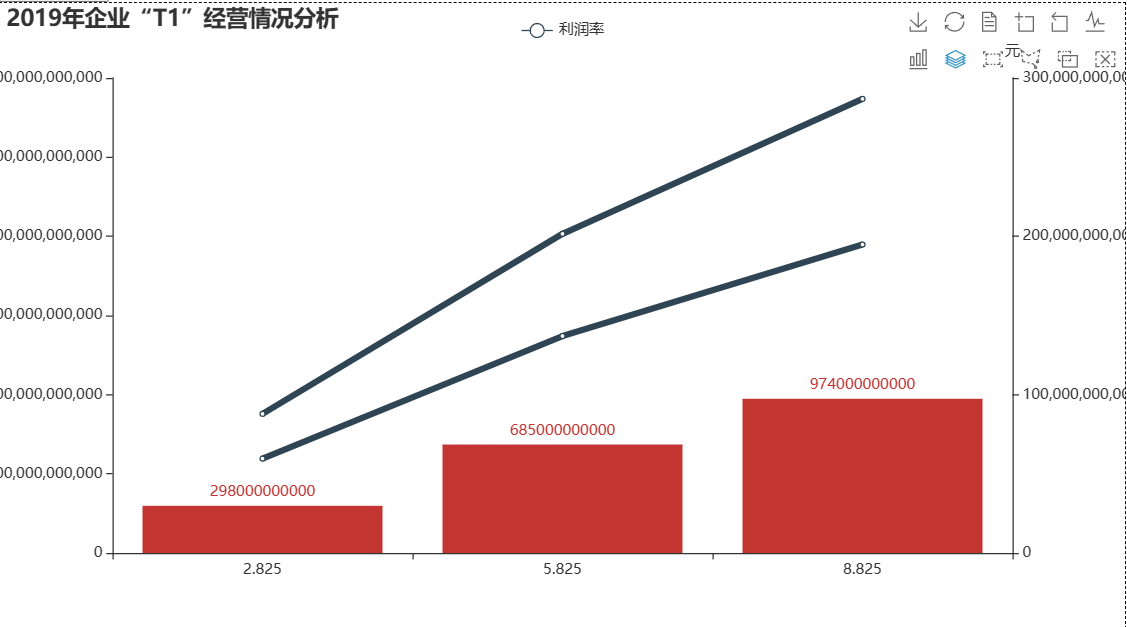
****

**图5 2019年企业601668的营业总成本分析**

如图所示营业成本为该企业第一大成本支出，其他部分相对偏少，故如需进行调整可以有限对营业成本进行调整可得到最高的效率和效果。

1. **绘制企业“T1”经营情况柱状折线组合图**

题中要求在同一张图表中，绘制企业“T1”2019 年3月、6月、9月三个季度营业总收入、营业总成本的柱状图和利润率、资产负债率变化的折线图，分析统计任务1处理后的数据，得出的柱状折线组合图如下图6所示：

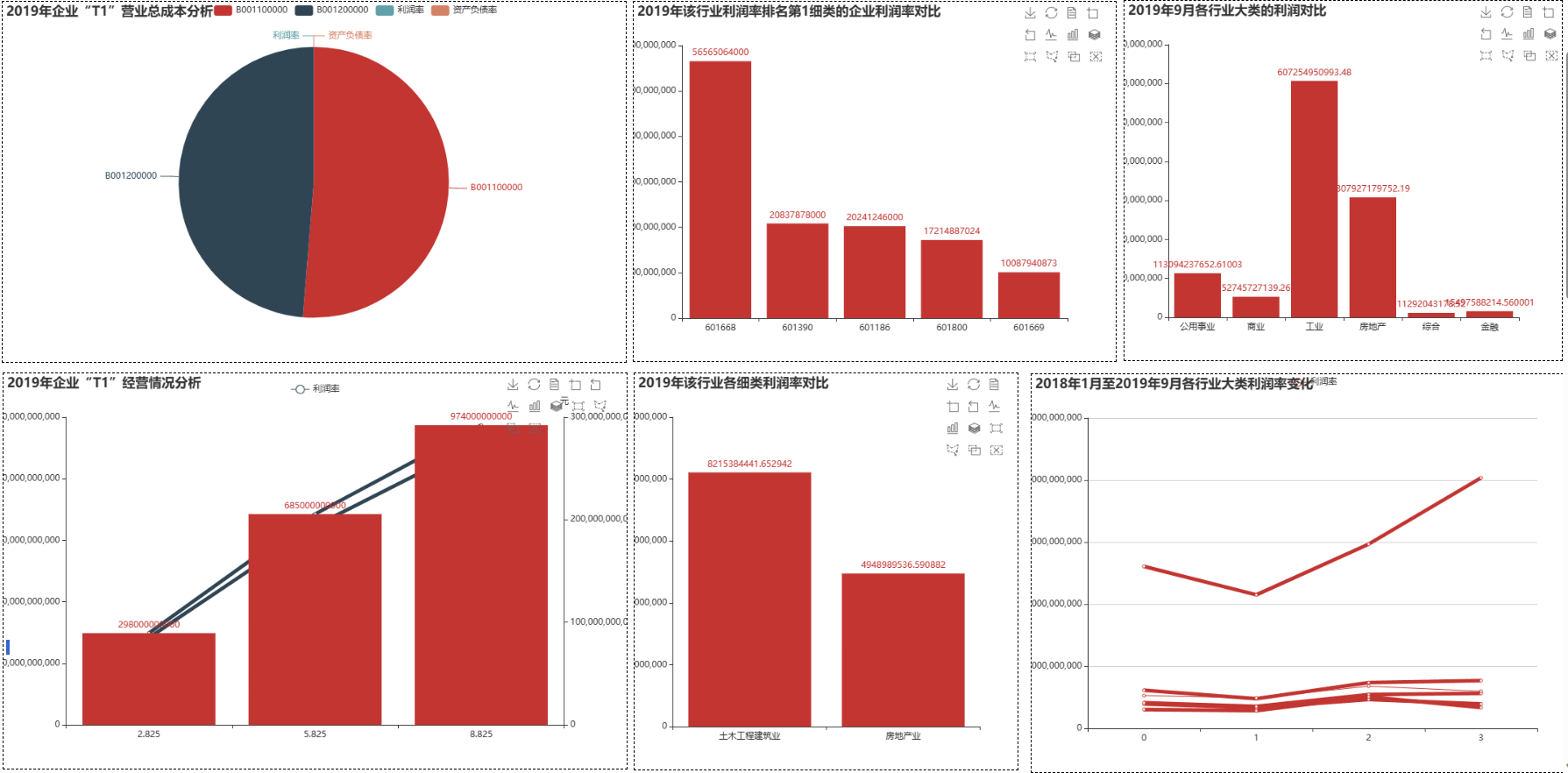
****

**图6 2019年企业601668经营情况分析**

如图所示601668季度呈现趋势性上涨，幅度大且有一定的平稳性，但其利润率与资产负载率相对平稳，无明显波动，但通过柱状图可看出当营业总收入增加时营业成本与收入差值越大，故整体来看利润率是呈现上升趋势。

**3.1.3任务2.3：行业与企业营业数据分析**

本问要求利用可视化大屏制作工具，将任务2.1和任务2.2所列的6张图制作成一个名为“行业与企业营业数据分析”。将可视化的图整合后得出图7如下所示：



**图7 行业与企业营业数据分析**

图7将6个可视化的图整合到一起，更加清晰直观看出各行业营业利润和利润率的对比及随时间的变化趋势，各行业与企业营收情况等。

1. **企业利润预测及财务造假识别**

**4.1任务3.1-3.3预测企业利润及识别财务造假**

在本问中，需要基于“”数据建立企业利润预测模型及财务造假识别模型。针对任务3.1，基于“”数据计算各个指标与利润总额的相关性，挑选相关度最高的5个指标。针对任务3.2，利用任务3.1挑选出的5个指标建立企业利润预测模型，运用建立的模型预测“”表中给定企业的利润总额。针对任务3.3，利用题目中的表6所列关键因子，对样本数据“”进行分析，建立财务造假的识别模型。并使用建立的识别模型对“”所列5个企业的财务数据进行筛查，识别其中唯一的1个涉嫌财务造假企业。

**4.1.1任务3.1：各指标与利润总额的相关性分析**

在本问中，基于“”数据使用  (皮尔森相关性系数）计算各个指标与利润总额的相关性，挑选相关度最高的5个指标。具体代码如下：

import numpy as np

import pandas as pd

from scipy import stats

data7 = pd.read\_csv("C:/Users/24453/Desktop/B题-企业财务数据分析与造假识别/数据/financial\_data.csv")

xgxs = data7.corr()

xgxs.to\_csv('C:/Users/24453/Desktop/B题-企业财务数据分析与造假识别/数据/相关系数矩阵.csv',index=False)

LRZE\_xgxs = data7.corr()["LRZE"]

LRZE\_xgxs.to\_csv('C:/Users/24453/Desktop/B题-企业财务数据分析与造假识别/数据/各个指标与利润总额的相关性.csv')

print(LRZE\_xgxs)

df4 = LRZE\_xgxs.sort\_values(axis=0, ascending=False)[1:6]

print(df4)

根据  (皮尔森相关性系数）计算各个指标与利润总额的相关度，得出最高的5个指标如下表5所示：

**表5 相关性最高的5个指标**

|  |  |
| --- | --- |
| 指标名称 |  |
| YYSR | 0.782726 |
| YWFY | 0.772832 |
| YYCB | 0.737736 |
| YYSJJFJ | 0.565440 |
| ZCJZSS | 0.238524 |

**4.1.2任务3.2：建立企业利润预测模型**

在本问中，利用任务3.1挑选出的5个指标建立企业利润预测模型，运用建立的模型预测“”表中给定企业的利润总额。

**（1）模型的构建**

一共建立了12个机器学习方法，分别为线性回归模型，回归模型，回归模型，岭回归模型，回归模型，多层感知机回归模型，决策树回归模型，极限树回归模型，随机森林回归模型，回归模型，梯度提升回归模型，回归模型。通过模型的建立并且对模型的测试值 和分别进行比较，可以得到越大，效果越好，正确率越高，具体结果见下表6所示：

**表6 不同机器学习对应的AUC值**

|  |  |  |
| --- | --- | --- |
| 机器学习方法 |  |  |
| 线性回归模型 | 0.919 | 0.874 |
| KNN回归模型 | 0.878 | 0.863 |
| SVM回归模型 | -0.021 | -0.03 |
| 岭回归模型 | 0.919 | 0.874 |
| LASSO回归模型 | 0.917 | 0.876 |
| 多层感知机回归模型 | 0.933 | 0.9 |
| 决策树回归模型 | 1 | 0.464 |
| 极限树回归模型 | 1 | 0.651 |
| 随机森林回归模型 | 0.972 | 0.852 |
| 回归模型 | 0.779 | 0.735 |
| 梯度提升回归模型 | 0.929 | 0.807 |
| Bagging回归模型 | 0.967 | 0.847 |

通过以上表格可以分析得到，决策树回归模型、极限树回归模型的训练值都为1，说明该方法过拟合，不采用。通过对其余模型的分析比较，最终选择使用多层感知机回归模型。

**（2）具体结果**

利用“”数据集对多层感知机回归模型进行训练，再运用建立的模型预测“”表中给定企业的利润总额。具体结果见下表7所示：

**表7 企业的利润总额预测结果**

|  |  |
| --- | --- |
| TICKER\_SYMBOL | LRZE |
| 4953174 | 455902781.7 |
| 4961537 | 645345799.3 |
| 4962538 | 187083429.6 |
| 4968740 | 168213601.9 |
| 4973917 | 97989526.2 |
| 4978589 | 522710981.2 |
| 4978721 | 149446068.6 |
| 4986535 | 152175714.5 |
| 4990739 | 290622224.7 |
| 4990942 | 122800470.7 |

**4.1.3任务3.3：财务造假识别**

在本问中，利用题目中的表6所列关键因子，对样本数据“”进行分析，建立财务造假的识别模型。并使用建立的识别模型对“”所列5个企业的财务数据进行筛查，识别其中唯一的1个涉嫌财务造假企业。

**（1）模型的构建**

一共建立了8个机器学习方法，分别为K-近邻模型，(高斯朴素贝叶斯)模型，(伯努利朴素贝叶斯)模型，决策树（信息熵）模型，决策树（系数）模型，神经网络模型，SVM模型，随机森林模型。通过模型的建立并且对模型的测试值****和****分别进行比较，可以得到AUC越大，效果越好，正确率越高，具体结果见下表8所示：

**表8 不同机器学习对应的AUC值**

|  |  |  |
| --- | --- | --- |
| 机器学习方法 |  |  |
| K-近邻模型 | 0.992 | 0.995 |
|  | 0.987 | 0.991 |
|  | 0.992 | 0.995 |
| 决策树（信息熵） | 1 | 0.994 |
| 决策树（系数） | 1 | 0.993 |
| 神经网络 | 1 | 0.993 |
| SVM | 0.995 | 0.994 |
| 随机森林 | 1 | 0.995 |

因为数据集数据大多FLAG属性为0，所以各分类模型拟合效果非常好，并不是出现过拟合现象。通过对其余模型的分析比较，最终选择使用随机森林模型。

**（2）具体结果**

利用“”数据集对随机森林模型进行训练，再运用建立的模型对“”表中5个企业的财务数据进行筛查，识别其中唯一的1个涉嫌财务造假企业。具体结果见下表9所示：

**表9 5个企业是否涉嫌财务造假结果**

|  |  |
| --- | --- |
| TICKER\_SYMBOL | FLAG |
| 4992858 | 0 |
| 4993201 | 0 |
| 4998808 | 0 |
| 4897311 | 1 |
| 4999709 | 0 |

如表所示，TICKER\_SYMBOL为4897311的企业涉嫌财务造假。