

复旦大学

期末项目报告

燃煤数据补缺及其经济意义挖掘

Author:

16307110288 张川
16302010063 郭涵青
16307090185 林雅文

2019 年春季学期“金融经济大数据挖掘”课程

in the

大数据学院

June 29, 2019

Contents

1 项目及数据介绍	1
1.1 项目背景	1
1.1.1 大数据与经济金融	1
1.1.2 燃煤数据补缺问题	1
1.2 数据概况	2
1.3 数据预处理	2
1.4 描述性分析	3
1.4.1 因变量：企业燃煤量	3
1.4.2 部门	4
1.4.3 隶属关系	4
1.4.4 登记注册类型	4
1.4.5 国有控股情况	6
1.4.6 行业大类	6
1.4.7 省份	6
1.4.8 成本相关：管理费用、财务费用、利息支出、产品销售成本、产品销售费用	6
1.4.9 资产相关：固定资产合计、实收资本、流动资产合计、所有者权益合计、资产总计	8
1.4.10 营收相关：出口交货值、存货、营业利润、应收帐款净额、工业销售产值、产成品、利润总额、总产值	8
1.4.11 负债相关：长期负债合计、流动负债合计、负债总计	8
1.4.12 税额相关：应交所得税、税金、本年应交增值税	10
1.5 总结与启发	10
2 模型部分	12
2.1 PCA	12
2.2 统计模型	12
2.2.1 模型结果	13
2.3 神经网络模型	15
2.3.1 分类变量预处理	15
2.3.2 数据去噪	16
2.3.3 分类神经网络模型	16

2.3.4 回归神经网络模型	16
2.3.5 数据填补	18
2.4 总结	18

List of Figures

1.1 变量说明	3
1.2 对数燃煤量分布情况	4
1.3 对数燃煤量与部门关系	4
1.4 对数燃煤量与隶属关系	5
1.5 对数燃煤量与登记注册类型	5
1.6 对数燃煤量与国有控股情况	6
1.7 对数燃煤量与行业大类	6
1.8 对数燃煤量与省份	7
1.9 对数燃煤量与成本相关指标的关系	7
1.10 对数燃煤量与资产相关指标的关系	8
1.11 对数燃煤量与营收相关指标的关系	9
1.12 对数燃煤量与负债相关指标的关系	10
1.13 对数燃煤量与税额相关指标的关系	10
1.14 连续变量之间的相关性矩阵图	11
2.1 主成分分析	12
2.2 统计回归模型结果	13
2.3 各模型预测值与真实值比较	14
2.4 聚类结果在前两个主成分上的分布	17
2.5 分类神经网络模型	17
2.6 回归神经网络模型	18

Chapter 1

项目及数据介绍

1.1 项目背景

1.1.1 大数据与经济金融

近年来大数据技术、人工智能技术及区块链技术等成为时代发展的潮流浪尖，这些技术在经济金融领域的应用也为人们所津津乐道。事实上，自经济学系统化发展成为一门学科以来，无论是理论研究还是计量模型，数据的搜集、处理和挖掘早就成为经济学研究的重要手段；数据之于建立在数字化市场之上的金融学更是如水之于鱼，不可或缺。但传统的数据处理手段往往局限于传统统计学手段，因此处理的问题往往不能太复杂，其所能应对的数据量也极其有限——特别在一些数据庞杂、变量间关系错综复杂的问题上，传统手段受到了很大挑战——而经济金融作为建立在大千世界上的社会学科，其研究的问题又往往具有这些复杂的特点。因此，我们可以发现新兴的大数据技术——这样一门为解决**巨大数据量、高维、非线性问题**而创立的学科——在经济金融学科上应当具有广阔的舞台。

在实际研究中，我们发现大数据分析技术可以用于处理大量数据的数据挖掘任务——例如日交易流水、淘宝交易清单等巨量数据；同时，例如机器学习、神经网络等方法又非常适用于在高维、高共线性、协整等传统方法表现极差的复杂关系数据集上的建模。概括的来说，**作为传统统计方法的复合、升级和延伸**，其在经济学研究中可以代替传统的统计学方法，帮助我们发现更多有用的信息。

1.1.2 燃煤数据补缺问题

在我国，工业企业是能源消费大户，其能源消费量往往能占到全国消费总量的 7 成以上。在传统重工业中，燃煤又是重要的能源来源，据浙江大学估计，仅 2017 年，全国工业锅炉燃煤消耗热量就达到约 5.4 亿吨。在微观上，燃煤量一定程度上反映出了这些企业的生产经营情况；在宏观上，通过比较不同行业、部门、省市间燃煤量的分布和变化，也可以帮助我们研究产业结构、产业布局、产业转移等经济的宏观情况。因此，研究工业企业燃煤量，对于我们研究工业企业具有重大意义。

研究燃煤量时，我们不仅需要研究单个变量的分布信息，还需要结合一些其他的企业特征。然而，在获取数据时，这些特征往往来自于不同的数据库，而不同数据库又经常存在不同的统计口径——例如工业企业数据库按照“企业销售规模是否超过 500/2000 万元人民币”进行入库选择，而工业企业污染物数据库则根据“是否属于重点污染物排放量前 85%”进行选择。这样不同的口径会导致数据集合上的困难——数据会产生大量空缺和偏差。这样直接合并的数据集存在一定样本偏误，如果直接进行使用可能会出现结论偏移。在此条件下，使用匹配上的真实样本训练相应预测模型，对未匹配上的数据进行相关数据预测是相对可行的方法。

我们的报告将着手解决企业燃煤数据的补缺问题，通过数据挖掘、机器学习和神经网络的方法，尝试不同的模型，选取出补缺效果最好的模型；同时，对于挖掘出的数据特征，我们希望应用经济学逻辑进行合理的解释。

1.2 数据概况

我们的数据来自于课程助教提供的数据集，其中包含 20000 个有燃煤量数据的样本和 30000 个无燃煤量数据的样本。因为解决的是燃煤量数据的补缺问题，我们利用了前 20000 条有标签的数据进行聚类分析、模型训练和交叉验证。

数据字段情况如下表1.1。其中除原有字段外，对于“行业类别”和“省地县码”两个字段我们做了提词处理，只保留大类信息，变成“大类行业类别”和“省份”两个字段。

我们可以看到，自变量中 ID 是重新从 1 开始编码的身份识别信息，除 ID 外，前几个变量都是分类变量且都用数字来表示，因此在放入模型之前需要预处理成分类变量（因子或字符串或哑变量）；后十几个变量都是连续的金额变量，而且基本上都是企业经营的财务数据，其中囊括了收入、利润、负债、税款等几个大的方面，每个方面包括了很多个指标，但这同时这也意味着这些数据间可能有很高的共线性、协整性，因此后续可能需要进行 PCA 降维或者变量筛选。

1.3 数据预处理

因为数据已经经过了一定的处理，因此我们的预处理只需要进行少量字段的再提取、异常值清洗和分类变量分离化处理。

所有字段中有两个分类字段“行业类别”和“省地县码”包含了多重信息，因为我们研究的是全国范围内各种行业的企业，数据分布较广，因此我们不需要也不希望一个变量有太多类。基于这个思路，我们进一步提取了“大类行业”和“省份”，以此减少分类数，方便后续建模。

变量类型	字段名	描述	补充说明
因变量	coalFuel	企业燃煤量	全部大于等于 0
自变量	ID	编号	共 50000 条, 用其中 20000 条有标注的
	sector	部门	分类变量
	belong	隶属关系	分类变量
	register	登记注册类型	分类变量
	share	国有控股情况	分类变量
	industry	大类行业类别	分类变量, 从行业类别中提取
	county	省份	分类变量, 从省地县 码中提取
	adminExp、asset、capital、 cash 、 equity 、 exportProduct 、 financeExp、incomeTax、 interest 、 inventory 、 longDebt 、 mainCost 、 operateExp 、 profit 、 receivable、saleProduct、 shortDebt、subInventory、 tax、totalAsset、totalDebt、 totalProduct、totalProfit、 valueAddedTax	管理费用、固定资产合计、 实收资本、流动资产合计、 所有者权益合计、出货交 货值、财务费用、应交所 得税、利息支出、存货、 长期负债合计、产品销售 成本、产品销售费用、营 业利润、应收帐款净额、 工业销售产值、流动负债 合计、产成品、税金、资 产总计、负债合计、总产 值、利润总额、本年应交 增值税	均为连续变量, 单位 为 : 千元

FIGURE 1.1: 字段表及变量说明

查看每个变量单独的分布情况, 会发现有的变量其实是存在一些异常值的, 这些异常值也会对后面的聚类、回归等等造成很大影响。具体的处理方法在每个模型的部分会详细介绍。

分类变量在原数据中是用数字表示, 直接放入模型会被当作连续数字。因此我们将他们处理成因子 (R 中) 或哑变量向量 (python 中), 包括一些语义嵌入的方法, 在后续模型中会更详细介绍。

1.4 描述性分析

1.4.1 因变量：企业燃煤量

如图1.2, 企业燃煤量有一些值为 0, 而另一些离群点值特别大, 因此做了其对数的分布图。

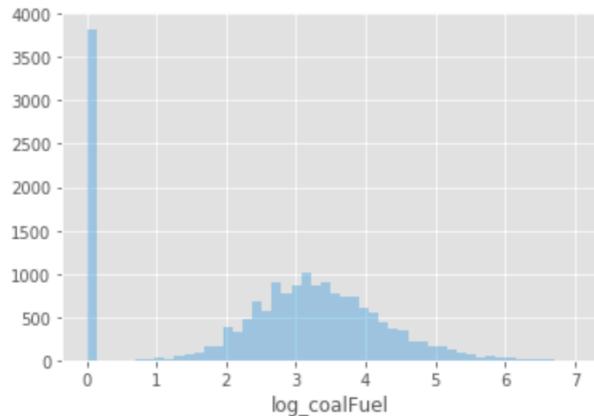


FIGURE 1.2: 对数燃煤量分布情况

· 分类变量

1.4.2 部门

如图1.3，不同部门间燃煤量差异明显，7、8 和 13 部门显著高出。

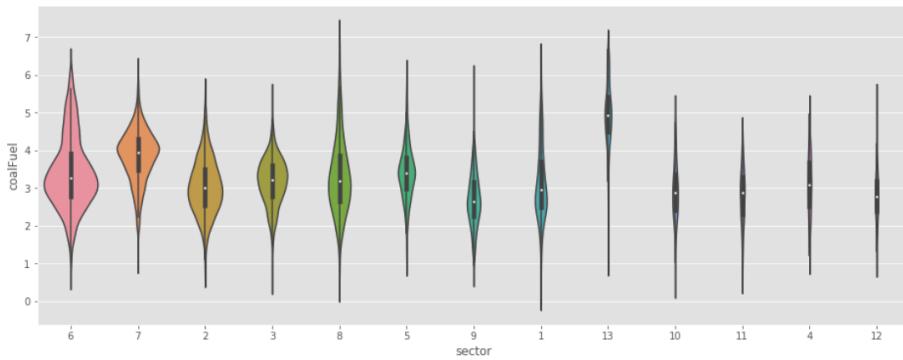


FIGURE 1.3: 对数燃煤量与部门关系

1.4.3 隶属关系

如图1.4，各隶属关系间燃煤量差异不大。

1.4.4 登记注册类型

如图1.5，151 和 340 显著高于其他类。

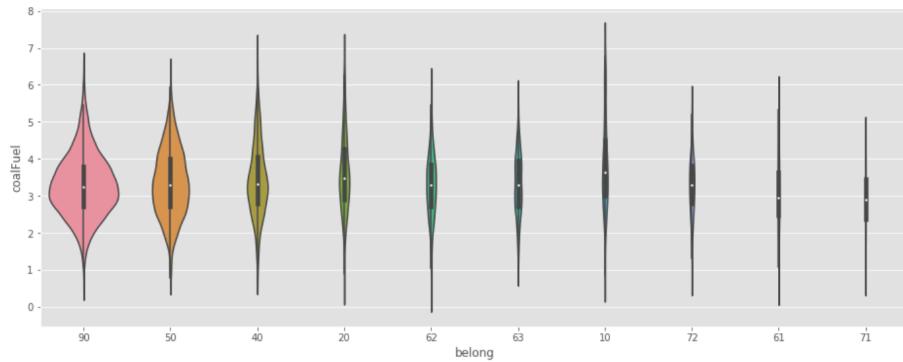


FIGURE 1.4: 对数燃煤量与隶属关系

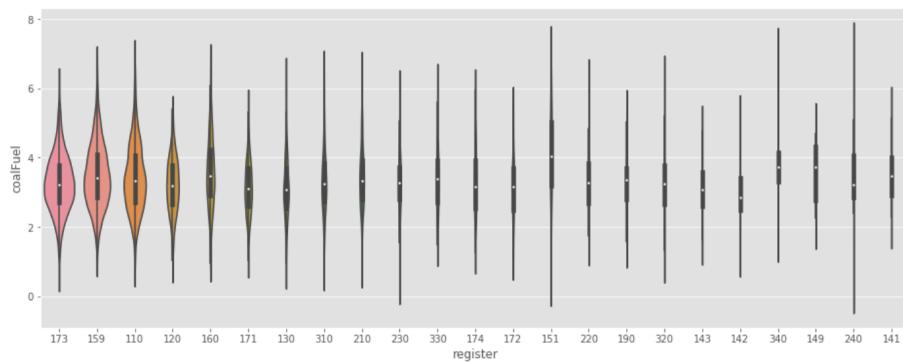


FIGURE 1.5: 对数燃煤量与登记注册类型

1.4.5 国有控股情况

如图1.6，除0类稍低一些外，差异不明显。

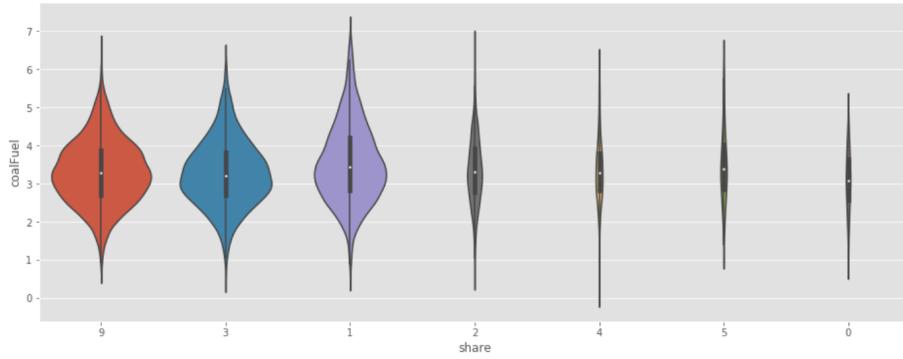


FIGURE 1.6: 对数燃煤量与国有控股情况

1.4.6 行业大类

如图1.7，32、44、25、28、45、12 和 2 类明显高于其他类别。

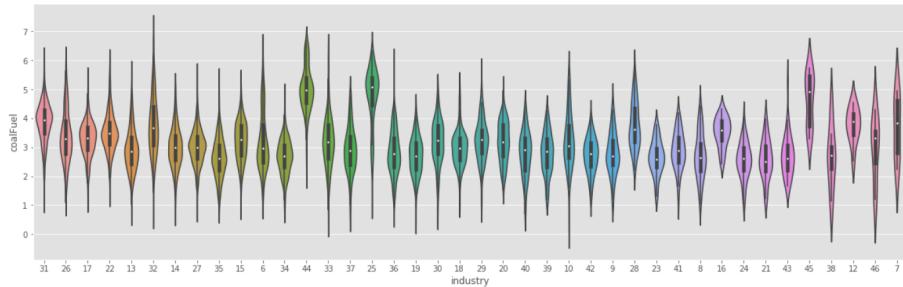


FIGURE 1.7: 对数燃煤量与行业大类

1.4.7 省份

如图1.8，14、53 和 64 这几个省普遍高于其他省。

- 连续变量

1.4.8 成本相关：管理费用、财务费用、利息支出、产品销售成本、产品销售费用

将成本相关的指标放在一起，查看它们各自的分布及其与燃煤量的散点图1.9。除产品销售成本外，其余指标与燃煤量关系都非常类似。

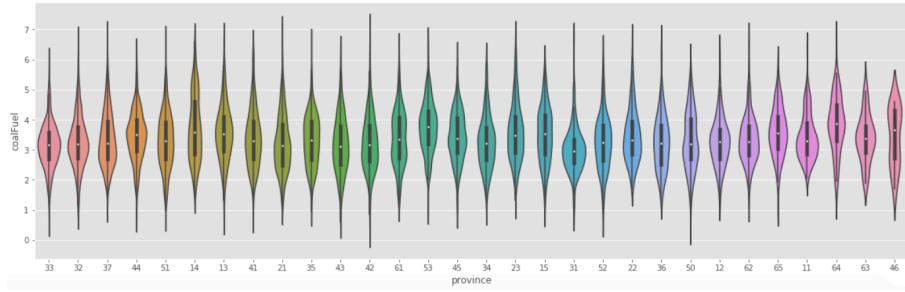
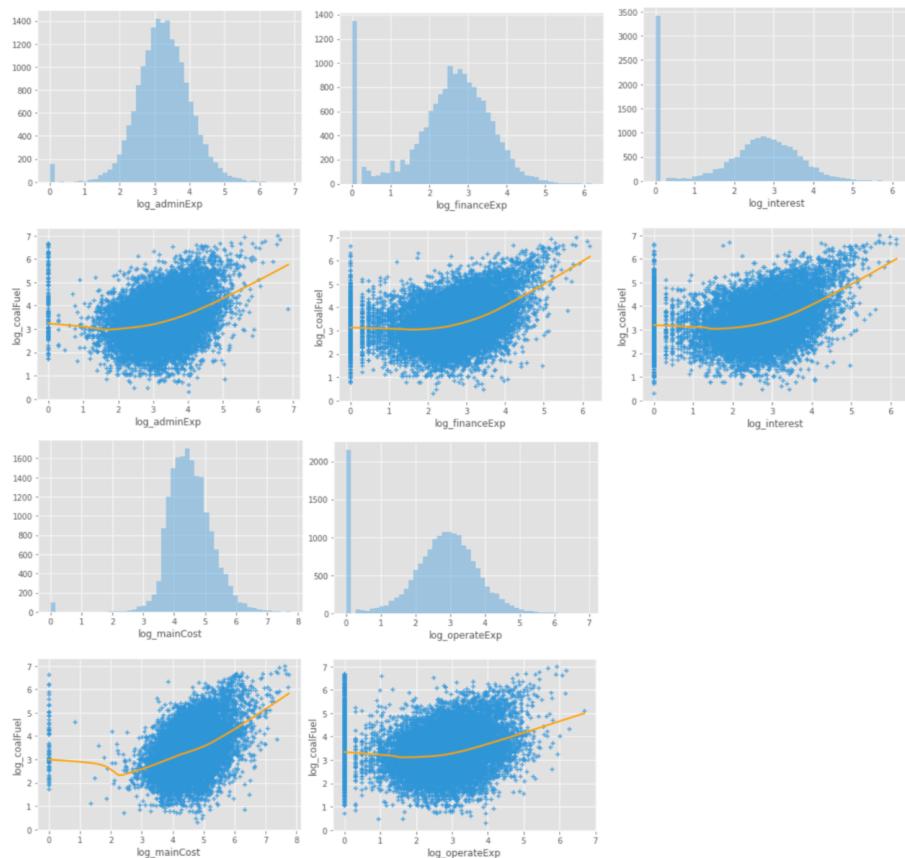


FIGURE 1.8: 对数燃煤量与省份



(1) 费用管理；(2) 财务费用；(3) 利息支出；(4) 产品销售成本；(5) 产品销售费用

FIGURE 1.9: 对数燃煤量与成本相关指标的关系

1.4.9 资产相关：固定资产合计、实收资本、流动资产合计、所有者权益合计、资产总计

将资产相关的指标放在一起，查看它们各自的分布及其与燃煤量的散点图1.10。几个指标与燃煤量的关系都非常类似。

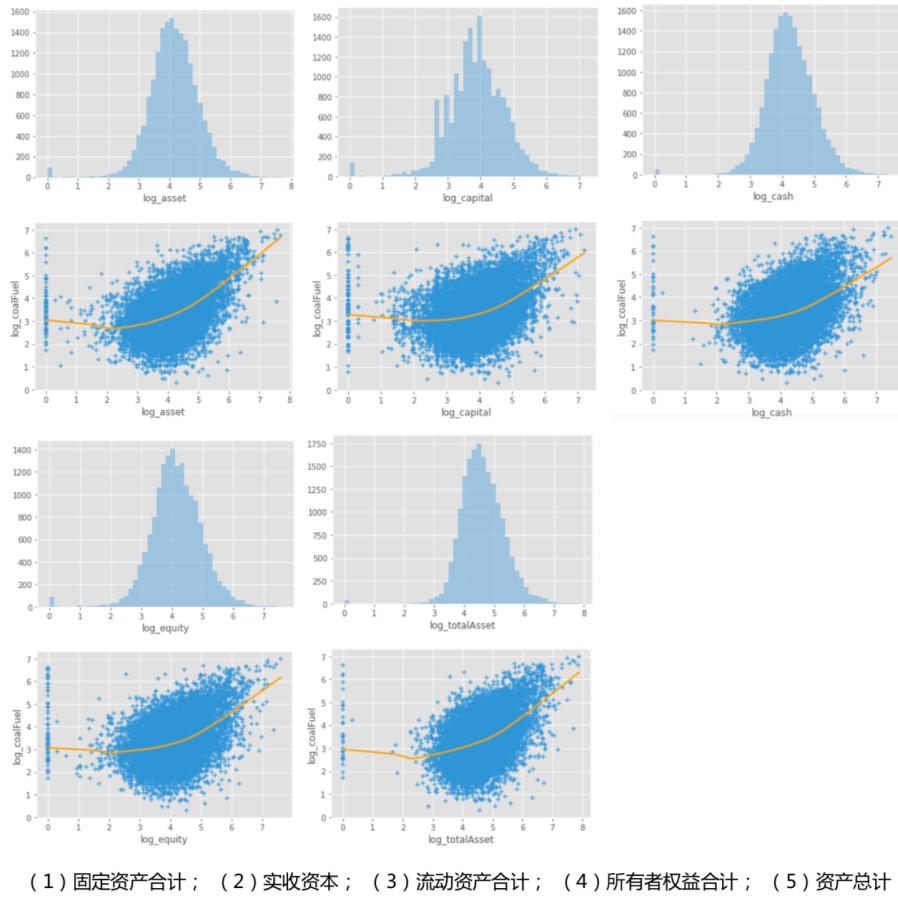


FIGURE 1.10: 对数燃煤量与资产相关指标的关系

1.4.10 营收相关：出口交货值、存货、营业利润、应收帐款净额、工业销售产值、产成品、利润总额、总产值

将营收相关的指标放在一起，查看它们各自的分布及其与燃煤量的散点图1.11。工业销售产值和总产值与燃煤量关系类似，而其余指标与燃煤量关系都非常类似。

1.4.11 负债相关：长期负债合计、流动负债合计、负债总计

将负债相关的指标放在一起，查看它们各自的分布及其与燃煤量的散点图1.12。几个指标与燃煤量的关系都非常类似。

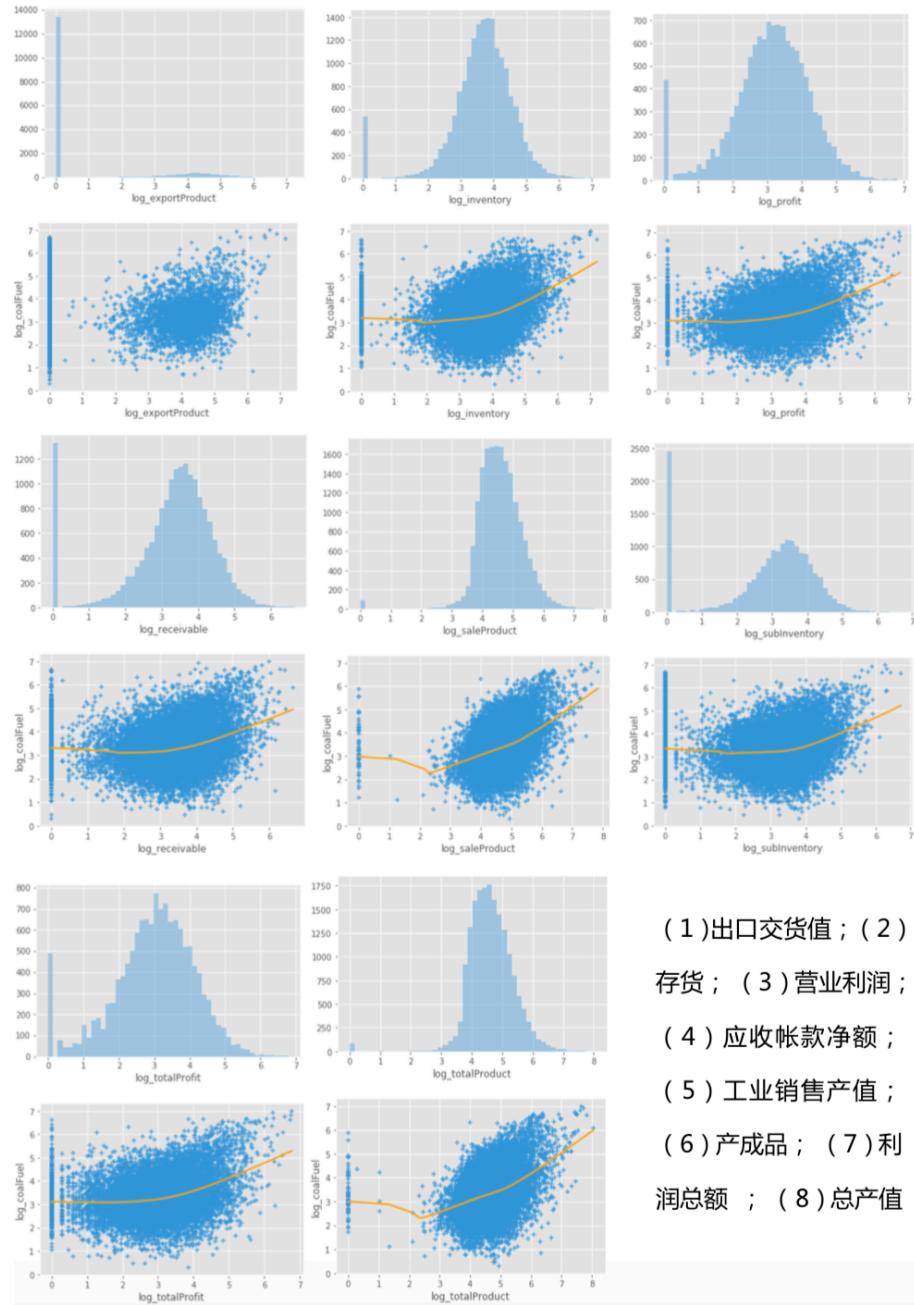


FIGURE 1.11: 对数燃煤量与营收相关指标的关系

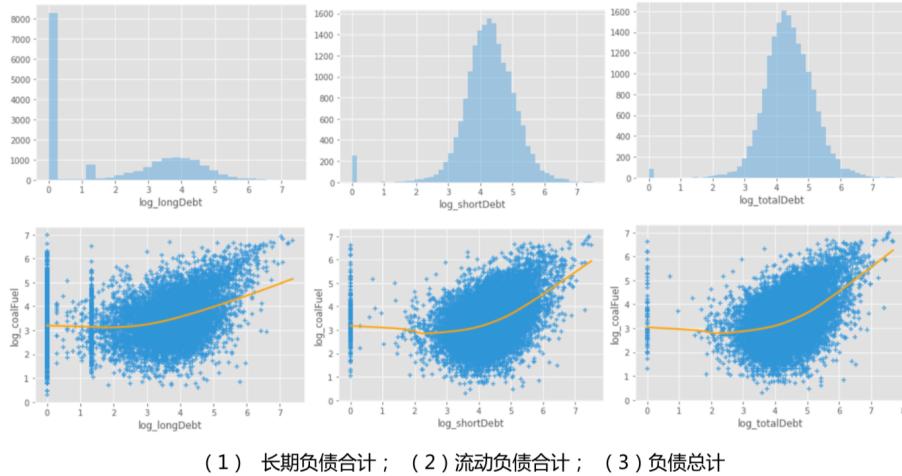


FIGURE 1.12: 对数燃煤量与负债相关指标的关系

1.4.12 税额相关：应交所得税、税金、本年应交增值税

将税额相关的指标放在一起，查看它们各自的分布及其与燃煤量的散点图1.13。几个指标与燃煤量的关系都非常类似。

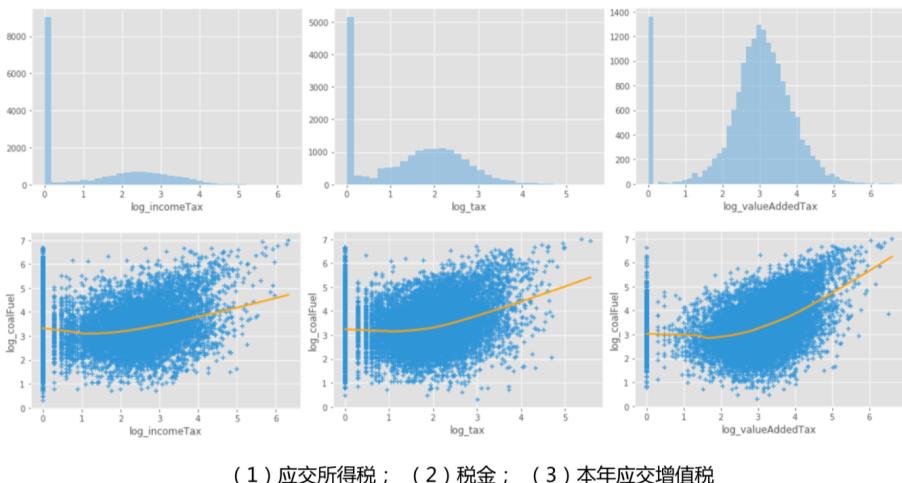


FIGURE 1.13: 对数燃煤量与税额相关指标的关系

1.5 总结与启发

从描述分析中，我们可以得到以下结论：

1. 分类变量中部门、登记注册类型、行业大类和省份对于燃煤量有较为显著的影响；
2. 连续变量中，几乎所有指标与燃煤量之间都有正相关关系；

3. 连续变量（财务数据）大致可以被分为成本、资产、营收、负债和税额五大类，而这五类中各包含多个指标，指标与指标之间、类与类之间都有很高的共线性（如图1.14）

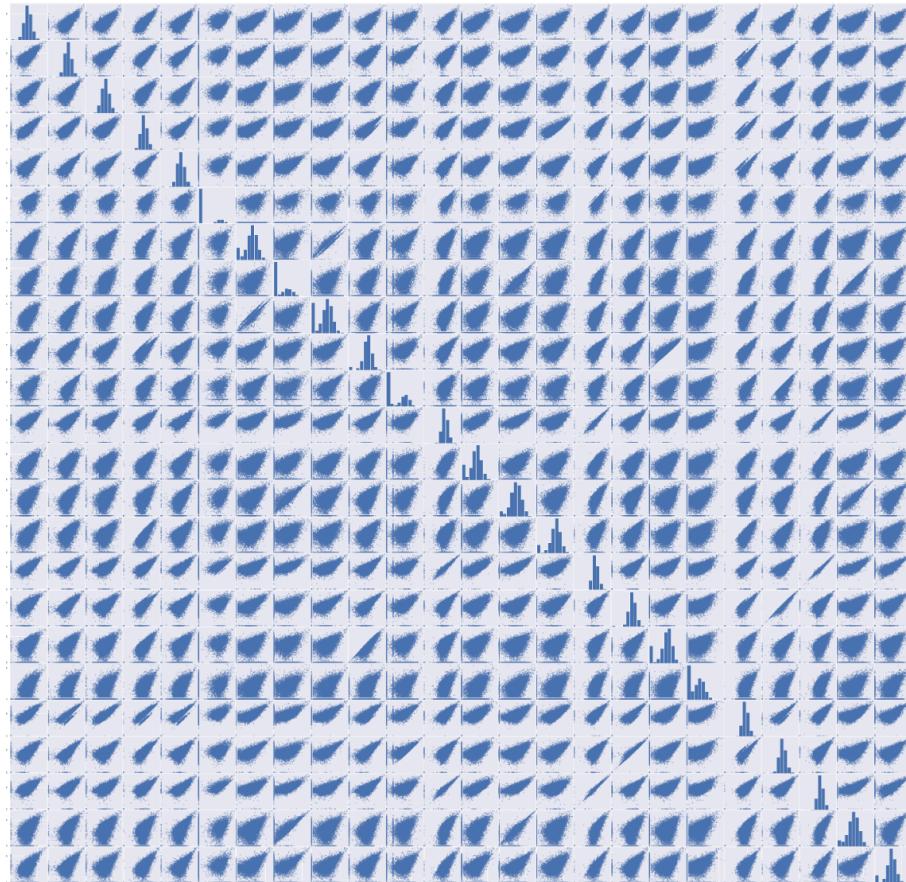


FIGURE 1.14: 连续变量之间的相关性矩阵图

Chapter 2

模型部分

2.1 PCA

考虑到连续变量间存在高度共线性，我们首先对所有连续变量做了主成分分析。根据每个主成分对于变异性的解释能力，我们发现前 5 个主成分已经可以解释 85% 以上的变异性，因此我们选取了各个样本在前五个主成分方向上的取值。各个变量在 5 个主成分方向上的取值分布如图2.1。

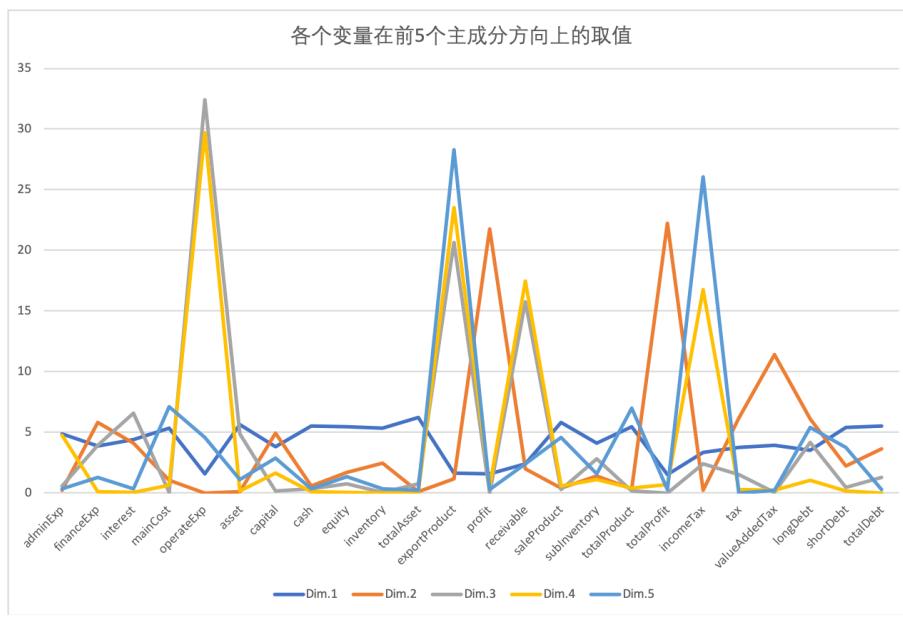


FIGURE 2.1: 主成分分析结果

2.2 统计模型

将编号 1-20000 为有标签数据按照 9:1 的比例划分为训练集、测试集，其中，两个数据集中两种标签比例相同：

1. 训练集：14175 个燃煤量大于 0 的数据，3825 个未使用燃煤的数据

2. 测试集：1575 个燃煤量大于 0 的数据，425 个未使用燃煤的数据

在统计建模过程中：

1. 对于模型结果我们使用 MSE 进行评估；
2. 分别在对数空间（以 10 为底对燃煤量进行对数变换）和原空间上进行训练；
3. 尝试使用 Ensemble 对多个模型的预测结果做加权平均；
4. 在原有变量的基础上，我们加入了两两之间的交互项。

2.2.1 模型结果

分别在燃煤量数据的原空间和其对数空间进行建模拟合，得到 MSE 如图2.2，预测结果与真实结果比较如图2.3。

模型	对数空间 MSE	原空间 MSE
Ridge Regression	9.25e+09	1.09e+10
LinearSVR	9.15e+09	-
KNN	1.41e+10	1.41e+10
RandomForest	9.22e+09	8.93e+09
GradientBoost	1.05e+10	8.12e+09

FIGURE 2.2: 统计回归模型结果

关于各模型，我们发现：

1. 线性模型：Ridge Regression 在对数空间上有更好的拟合效果。加入交互项后，该模型的性能有显著提升。由于 LinearSVR 在原空间中存在数值问题，训练结果异常，因此我们不在原空间中训练该模型；
2. 树模型：随机森林和 GradientBoost 均在原空间上有更好的拟合效果。其中 GradientBoost 的 MSE 最低；
3. K 近邻模型：从拟合效果图可以看出，KNN 同样无法将燃煤量为 0 的数据预测出来，说明燃煤量为 0 的数据与其他数据没有很好的区分度，因此尽管可以尝试做分段预测，但很可能对 MSE 不会有太大的提升；
4. 在原空间与对数空间上的拟合效果对比：由于评价指标为 MSE，因此各模型均在原空间上拟合的 MSE 更小，但从原空间上的拟合效果图，我们可以发现总体上拟合的效果并不佳，主要存在以下问题：

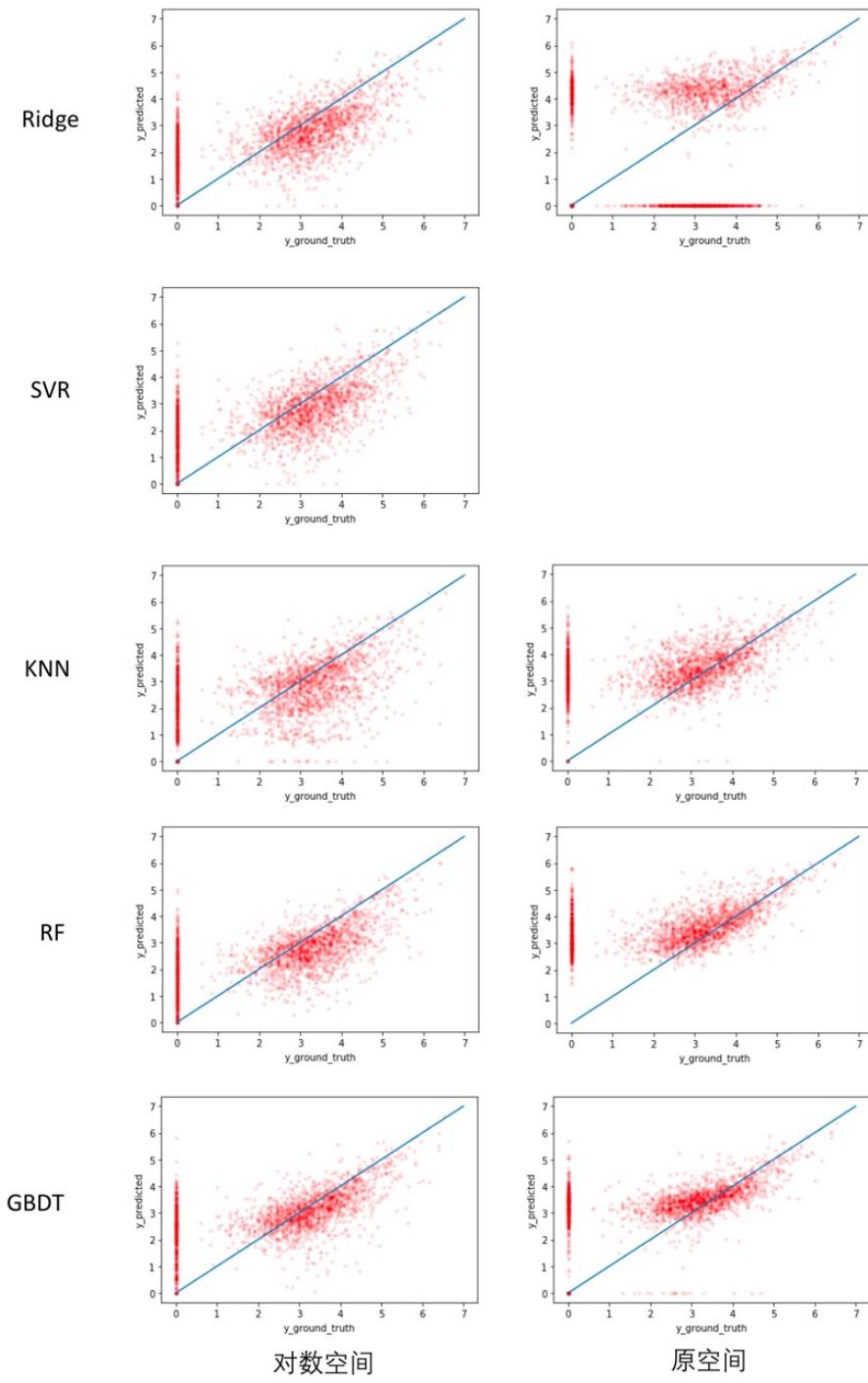


FIGURE 2.3: 各模型预测值与真实值比较

- 1) 燃煤量真实值较小的数据，预测值普遍偏大；
- 2) Ridge Regression 和 GradientBoost 预测出来的结果会有负值。

而从对数空间上的拟合效果图，我们发现总体上对数空间上拟合的效果更好，对于燃煤量真实值较小的数据有更逼近真实值的预测结果，但由于评价指标为 MSE，因此在对数空间上拟合的 MSE 普遍偏大，只有线性模型 Ridge Regression 在对数空间上的 MSE 小于原空间；

5. 总体上，GradientBoost 模型的预测效果最佳，但各模型都无法准确预测燃煤量为 0 的数据；
6. 最后尝试了将上述模型进行集成，但最终的结果都不如直接使用 GradientBoost 来得好。

如前面分析中提到的，燃煤量为 0 的样本带来了很大的 MSE，我们可以考虑先用分类模型区分开燃煤量为 0 和不为 0 的样本，再在不为 0 的样本上训练回归模型。但是实验后我们发现，用传统统计方法得到的分类效果不佳，如随机森林模型进行分类，模型 AUC 能达到 80%，但是用此模型先预测，其将有燃煤量的企业预测为无燃煤量带来的误差也非常大，因此用统计模型先分类再回归对于最终 MSE 并没有提升。

2.3 神经网络模型

人工神经网络是一种模拟动物神经网络行为特征进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，达到处理信息的目的。人工神经网络有以下特点：

1. 知识以分布方式存储在整个系统；
2. 具有很强的容错能力；
3. 理论上已经被证明可以逼近任意复杂的非线性系统；
4. 具有良好的自适应、联想等智能，能适应系统复杂多变的动态特性。

正是以上特点，使其在变形监测数据处理与分析上有着独特的优越性。在我们的数据中存在大量的共线性关系，因此用传统统计模型难以很好地表达出模型的非线性。因此我们考虑用神经网络模型来进行预测和分类。

2.3.1 分类变量预处理

预处理主要将分类变量转换为连续变量，用到了分词和 word2vec 方法，为后续回归模型训练做准备。

先将分类变量处理为 dummies (哑变量)，共有 6 个分类变量，每个分类变量值域个数不同，总计共有 124 个 dummy 变量。若某一行 sector 的值为 2，则该行 sector_2 属性的值为 1，以此为所有行生成属性变量。因为一行数据可能具有多个属性，因此可以看成一句句子，如 ['sector_1', 'belong_2', ...]。我们为每一行都创建一句句子。因为 dummy 变量个数太多，所以我们训练了一个 word embedding 模型来进行降维，使得每一个属性都能用 32 维的单词向量来表示。最后，我们把一句句子中的每个单词向量求和取平均，就是这行数据的句子向量。

至此，我们顺利地把 8 个分类变量转换为了 32 个连续变量，可以用数值距离代表数据之间的差异。以后提到的数据，都默认将分类变量转换为了连续变量。

2.3.2 数据去噪

将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。聚类本身就是最常用的异常值检测方法，大部分非监督的异常值检测都依靠聚类，例如 K-means 对离群值非常敏感。

我们在所有数据上进行聚类（不看 ID 和 coalFuel），以去除噪声值。聚类后的结果用 PCA 进行降维，在前两个主成分方向上展示如图2.4。我们可以看到 3 类分布的差异性非常明显，因此可以轻易判断出蓝色和绿色的两类为噪声。因此我们选取红色的类，在其上进行后续的模型训练。以后提到的数据，都默认剔除了噪声值。

2.3.3 分类神经网络模型

如果能够判断一个公司是不是使用燃煤生产，之后只在使用燃煤生产的公司进行回归模型预测，则结果会更准确。因此我们以有无使用 coalFuel 为因变量 ($\text{coalFuel} > 0$ 则为 1, $= 0$ 则为 0)，除 ID 之外的属性为自变量，在给了 coalFuel 的训练数据上训练神经网络分类模型。训练过程如图代码2.5。

最终的分类器在测试集上的分类准确度达到了 86%，效果相较传统统计方法提升不大。

2.3.4 回归神经网络模型

同样，我们使用神经网络回归器来进行回归模型训练。

我们的神经网络模型在剔除 coalFuel 为 0 的训练数据上进行训练，设置参数为：迭代 5000 次，三层隐单元层，adam(鲁棒性较好) 方法。代码如图2.6。

最后的 MSE 为 $6e9$ ，比最好的统计模型误差优化了近 $1/3$ 。

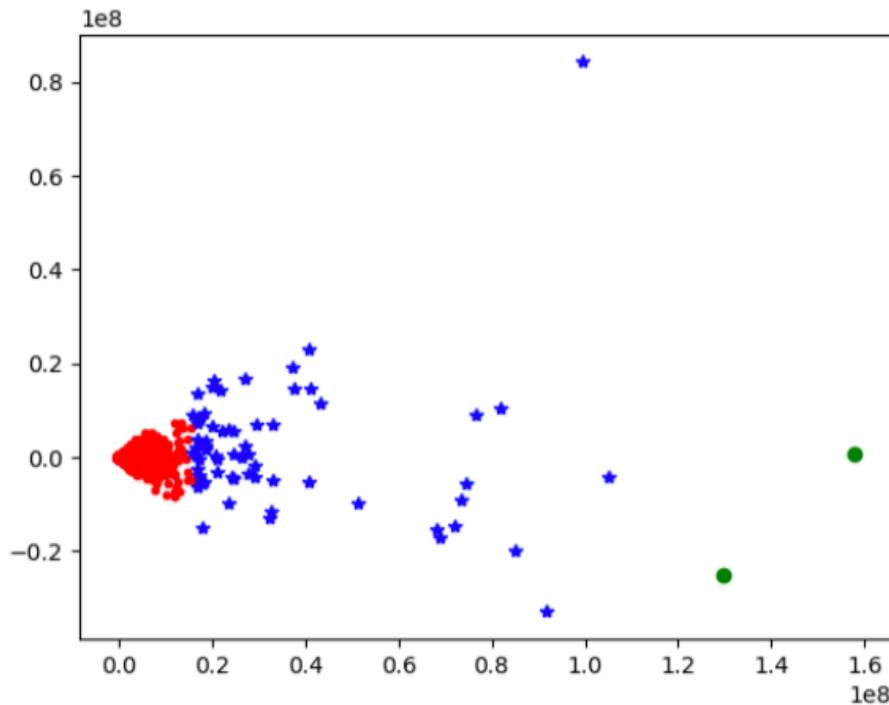


FIGURE 2.4: 聚类结果在前两个主成分上的分布

```
test_i = 15900
train_x = df.iloc[:test_i, 2:-1].values
train_x = scaler(train_x)
test_x = df.iloc[test_i:19935, 2:-1].values
test_x = scaler(test_x)
train_y = df.iloc[:test_i, -1].values
test_y = df.iloc[test_i:19935, -1].values

fname = 'svm.model'
if os.path.exists(fname):
    clf = joblib.load(fname)
else:
    clf = MLPClassifier(solver='adam', max_iter=2000, alpha=1e-5,
hidden_layer_sizes=(5, 2))
    clf.fit(train_x, train_y)
    joblib.dump(clf,fname)
predict = clf.predict(test_x)
MSE = np.sum(np.power((test_y - predict), 2))/len(test_y)
print(MSE)
```

FIGURE 2.5: 分类神经网络模型

```
def regress(fname, num, iter):

    df = pd.read_csv(fname, header=0, encoding='utf-8')
    df = df.iloc[:num, :]
    df = df.loc[df.coalFuel != 0]

    train_x = df.iloc[:num, 2:].values
    train_x = scaler(train_x)
    train_y = df.iloc[:num, 1]

    fname = 'regress.model'
    if os.path.exists(fname):
        clf = joblib.load(fname)
    else:
        clf = MLPRegressor(solver='adam', max_iter=iter, alpha=1e-5,
hidden_layer_sizes=(5, 3, 3))
        clf.fit(train_x, train_y)
        joblib.dump(clf, "regress.model")

    predict = clf.predict(train_x)
    print(predict.tolist())
    MSE = np.sum(np.power((train_y - predict), 2))/len(train_y)
    print(MSE)
    return clf
```

FIGURE 2.6: 回归神经网络模型

2.3.5 数据填补

最后，我们使用分类和回归两个神经网络模型在 50000 个数据的全集上进行预测，预测是否使用燃煤量以及使用的多少。

其中，我们预测出 50000 个公司中共有 3274 个不使用燃煤，其余预测结果以 csv 文件格式保存在”predict/result.csv”文件中。

2.4 总结

从以上建模结果，我们可以发现在传统统计模型中，GardientBoosting 在回归问题上处理效果最好，但其 MSE 依然达到 8e9；而相比之下，我们使用并不复杂的分类 + 回归神经网络模型，就可以将最后的预测误差 MSE 降到 6e9，比之前的结果优化近 1/3。但同时，神经网络也存在结果不够稳定的问题。

然而，即使使用了神经网络模型，该 MSE 依然在一个较大的数量级上，也就是说平均下来在每个样本上预测误差（RMSE）在 77000 左右，这个值依然是很大的。对于

神经网络模型来说，我们目前的数据量并不算很大，因此如果拥有更多训练样本，相信我们的神经网络模型可以取得更好的预测效果。

综上，虽然统计模型和神经网络模型在数据填补上表现还有提升的空间，但是其对于我们校正数据依然具有指导作用，尤其是在样本量不断积累的情况下，相信神经网络模型能有更进一步的表现。