## 数据来源

**（1）图像数据集：**由遥感数据由具有较高的空间分辨率和时间分辨率的哨兵二号卫星得到，已经过气象处理，再进行云层处理得到分辨率为10m的农作物图像。

**（2）预测数据集：**通过民用卫星哨兵二号获取多光谱图像(GEE)，选用其中的第8波段为红外波段，第4波段为红光波段，进行两者之差比两者之和，提取农作物四个生长时期的NDVI，进行生长态势分析，进行生长态势的评估作为预测的特征向量

**（3）气象：**国家青藏高原科学数据中心

**（4）化肥：**农业知识服务系统搜索

**（5）亩产量：**农业知识服务系统搜索

## 2. 数据规模

预测数据量为1872，是一个1872行，14列的二维数组。

图像数据为200个农作物图像及其标签，以及为扩大数据集后续对图像进行的增广。

## 3.开发工具与技术

依照上述陈述，本作品分为四个部分：Python后端、SSM后端、前端、Jupyter 分析文件。其中 Jupyter 分析文件是进行数据处理与数据分析的文件，不需要部署，其他三项都需部署。

Python 后端需在 Python3.7及以上环境运行。通过安装最新版 Anaconda，并进入到项目目录下，使用pip工具安装requirements.txt中的依赖包。随后，使用Python分别运行Forrest.py以及FCN.py，从而启动图像识别以及预测服务。由于使用了深度学习技术进行高精确的识别和预测，对机器的内存、CPU 要求较高，因此部署 Python 服务时务必确保机器中至少有4GB可用内存。

SSM后端采用IDEA2021个人版进行开发。安配置好NodeJS和tomcat后，进入项目目录，执行npm install 命令安装依赖，随后执行命令 node bin/www 启动后端服务。

前端采用VUE技术打包，需要安装NodeJS。安装后，进入项目目录，执行npm install命令安装依赖，随后执行命令npm start启动前端服务器，使用浏览器访问本地8080端口浏览页面。后端依赖Redis、MySQL等数据库，数据的安装与部署不再赘述。

## 4.测试过程

### 4.1 后端测试报告

表5-1 测试报告

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **测试项目** | **测试内容** | **测试要求** | **测试结果** | **备注** |
| 软件结构测试 | 是否使用B/S架构 | 1.判断客户端输入有效  2.判断服务器可以对客户端的输入做出正确的响应  3.关闭服务后客户端输入无效 | 使用了 B/S 架构对客户端属于做出正确响应 |  |
| 网络结构测试 | 验证网络数据流向 | 使用 TCPDUMP 验证节点间数据流向 | 网络数据流向正常 |  |
| 先进性测试 | 根据系统总体结构和采用的技术对系统的先进性作出评价 | 由相关组织，部门，机构或个人在了解系统构成的基础上对系统的先进性作出合理评判 | 系统设计模式为MVC 设计模式 |  |
| 开放性测试 | 是否采用标准协议，接口规范 | 由相关组织，部门，机构或个人在了解系统构成的基础上对系统的开放性作出合理评判 | 符合标准协议，接口规范 |  |

### 4.2 数据处理与测试

**（1）图像数据集增广**

图像增广在对训练图像进行一系列的随机变化之后，生成相似但不同的训练样本，从而扩大了训练集的规模。 此外，应用图像增广的原因是，随机改变训练样本可以减少模型对某些属性的依赖，从而提高模型的泛化能力。 例如，我们可以以不同的方式裁剪图像，使感兴趣的对象出现在不同的位置，减少模型对于对象出现位置的依赖。 我们还可以调整亮度、颜色等因素来降低模型对颜色的敏感度。具体处理方式如下图：

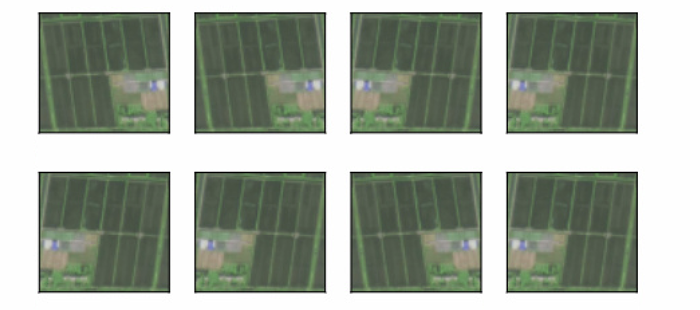


图5-1 不同方位增广

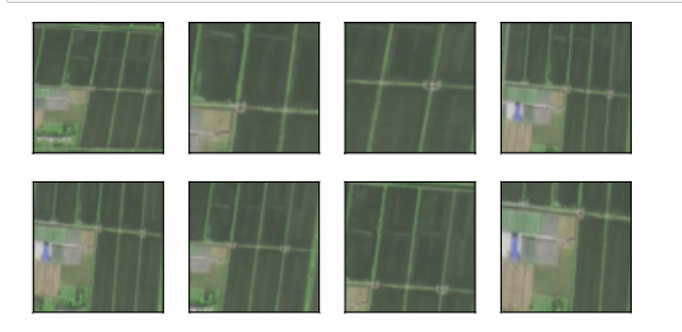


图5-2 进行缩放增广

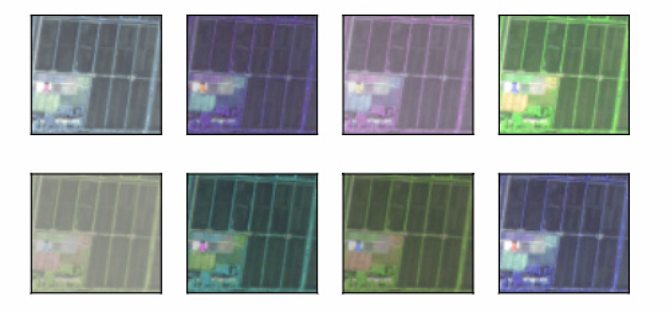


图5-3 颜色与亮度调节进行增广

1. **相关数据处理**

首先是导入数据，进行查看。如图5-4所示，可以看到有13个特征值，1个目标值，数据量为1872，是一个1872行，14列的二维数组。

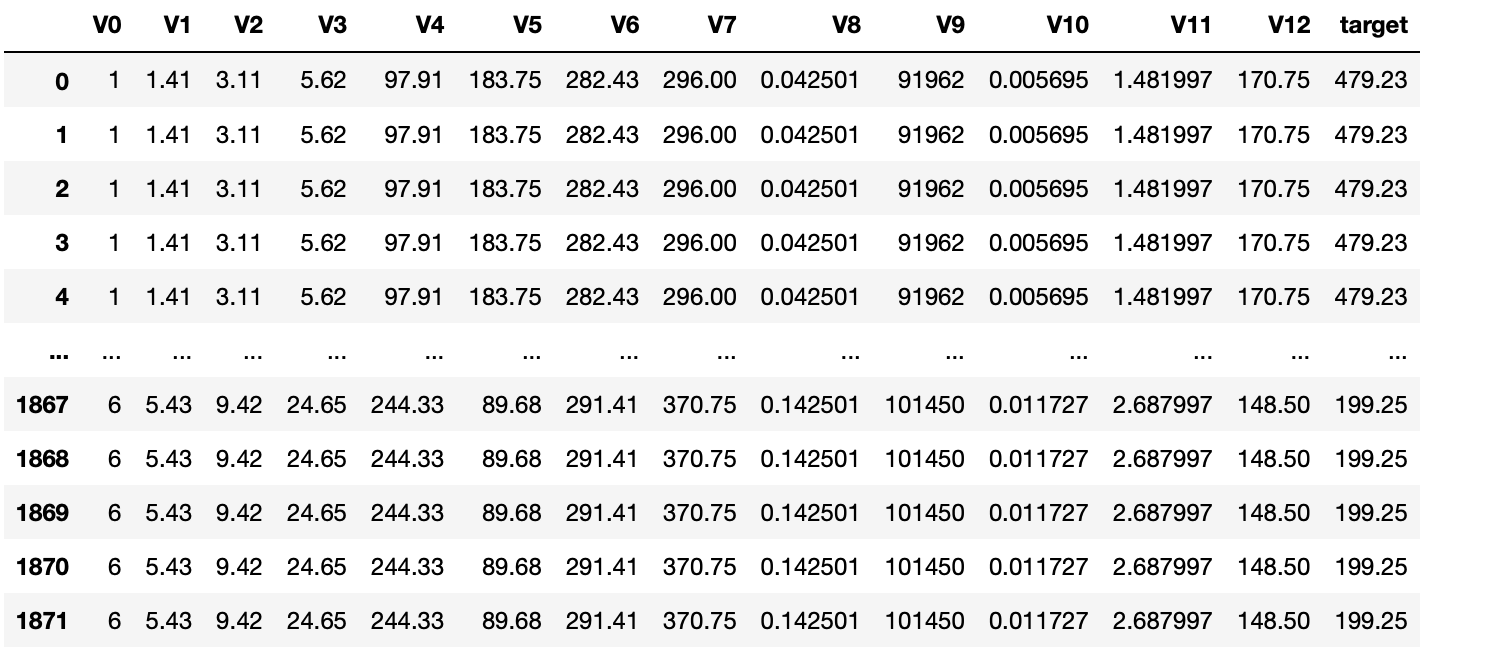


图5-4 数据集导入

这里使用Robust方法对数据集进行归一化处理，可以使是模型具有较高的精度或有效性，这也是对于机器学习中所有学习模型的基本要求。去除噪声和离群点的影响，因为用鲁棒性理解述算法对数据变化的容忍度有多高。鲁棒性并不同于稳定性，稳定性通常意味着特性随时间不变化的能力，鲁棒性则常被用来描述可以面对复杂适应系统的能力，需要更全面的对系统进行考虑。处理结果如图5-5所示。

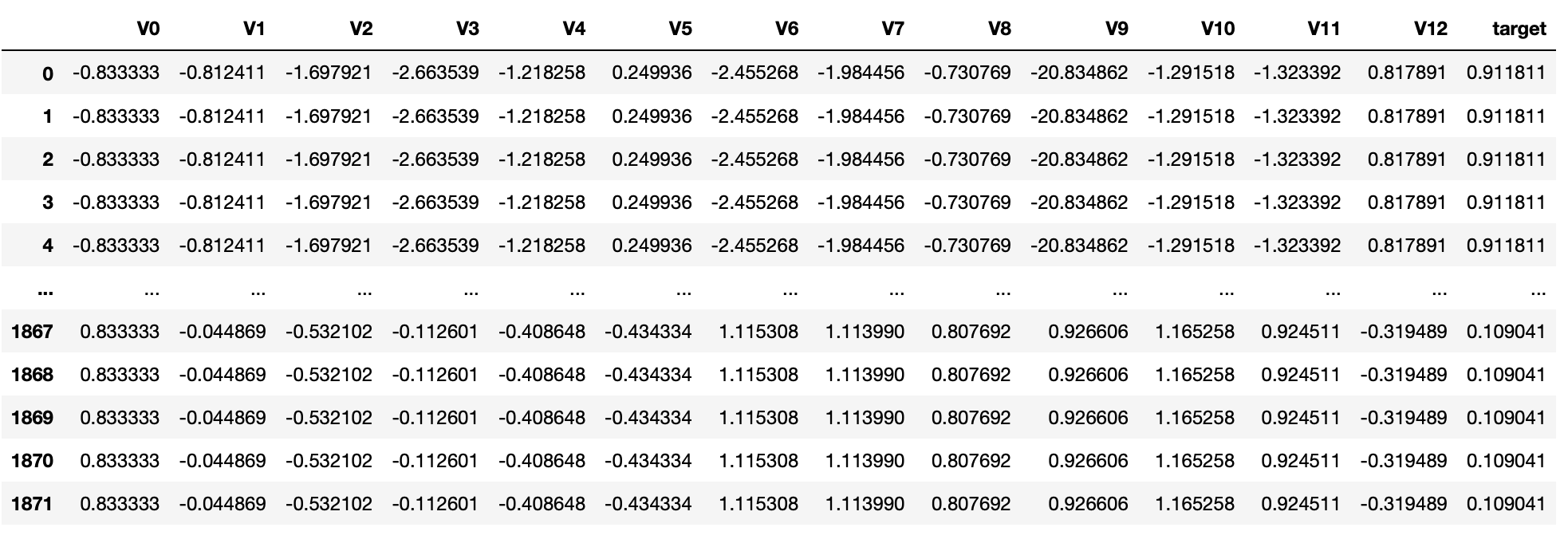


图5-5 归一化处理

接下来查看箱式图，可用于异常值检测

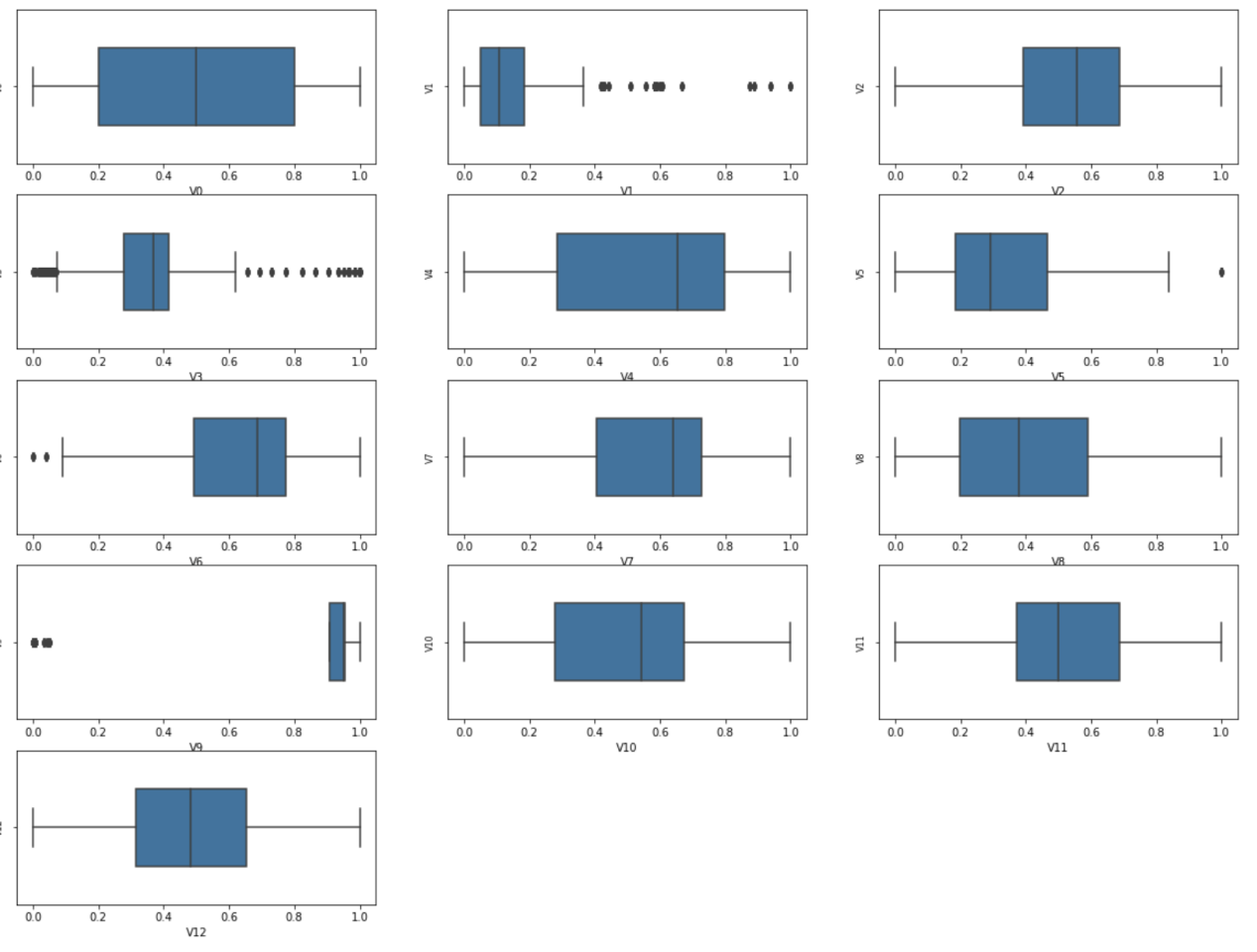


图5-6 箱式图

观察图5-6可以发现V1,V3有很多数据偏离正常值，处于四分位点以下，故这些数据存在许多较大的异常值，可以移除。

查看特征变量的数据分布直方图，并绘制Q-Q图查看数据是否近似于正态分布。

这是指数据的分位数和对比参照的图，如果数据符合正态分布则所有的点都会落在直线上。

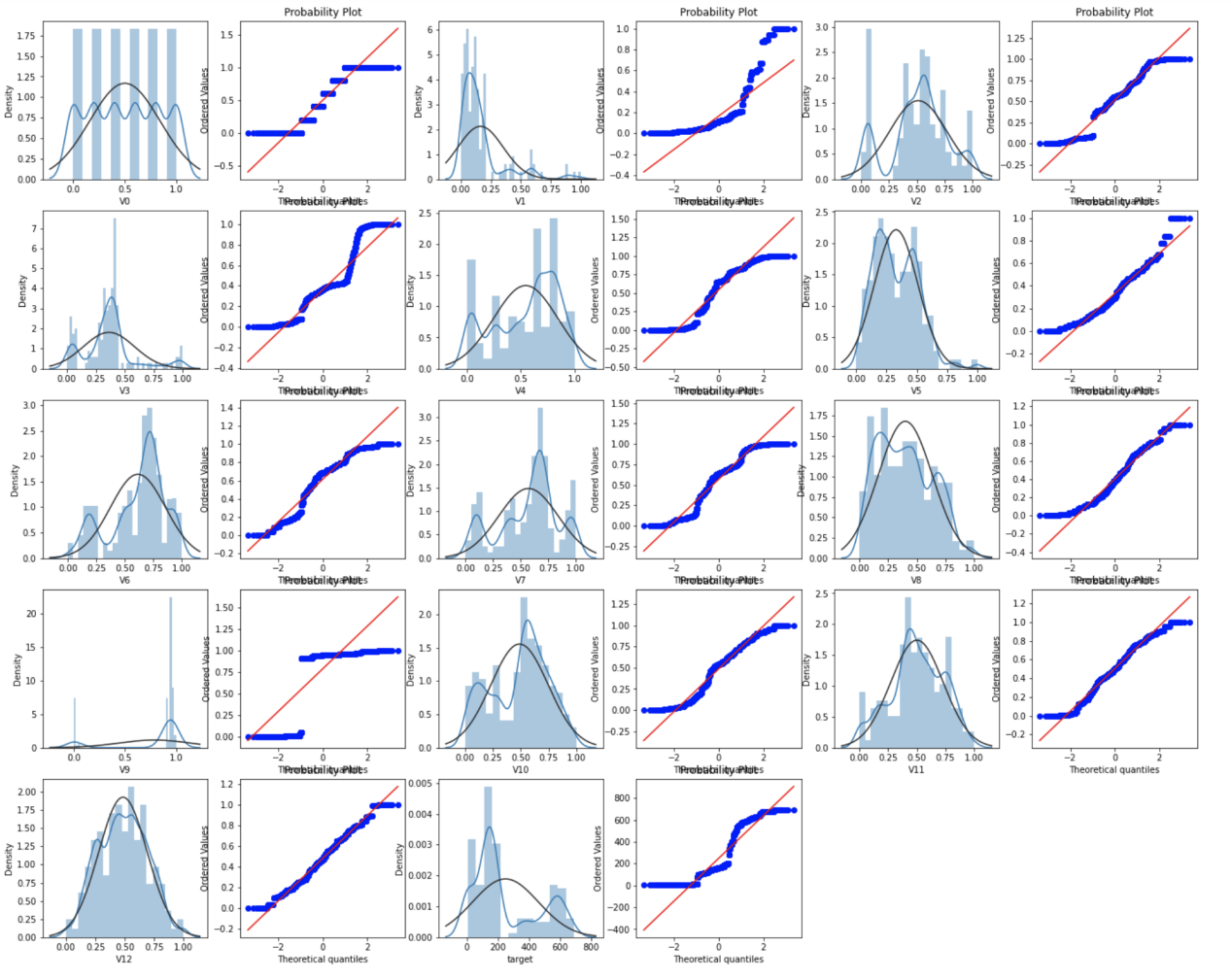


图5-7特征变量的数据分布直方图

如图5-7可以发现部分数据不随对角线分布，后续可以使用数据变换。

接下来对数据的相关性分析，可以发现特征变量和目标变量及特征变量之间的关系，为提取特征做准备，画出相关热力图，结果为所有特征变量和target变量之间的相关性系数，对角线数据一样。

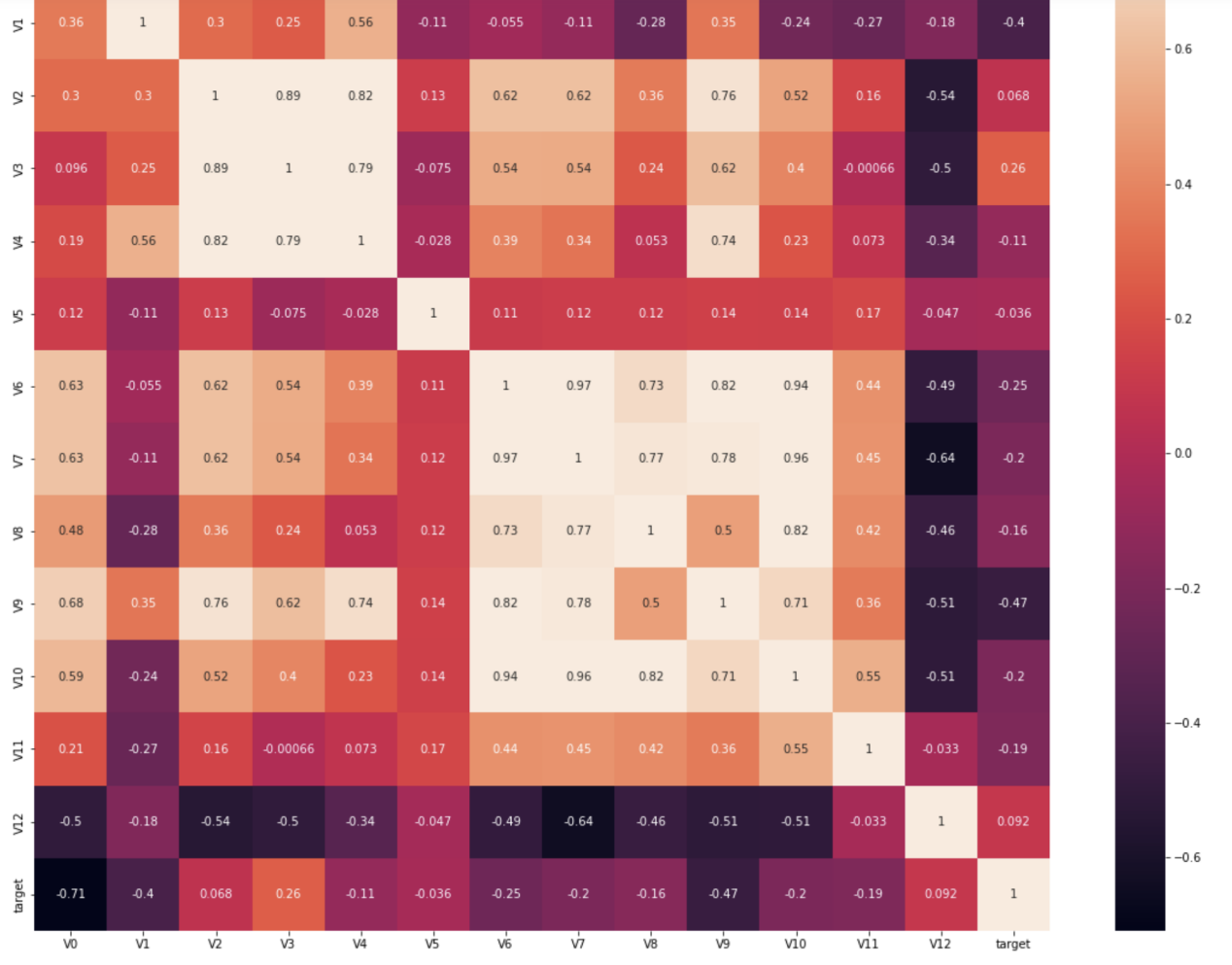


图5-8 热力图

只保留梯形热力图，如图5-9所示，找出相关度。

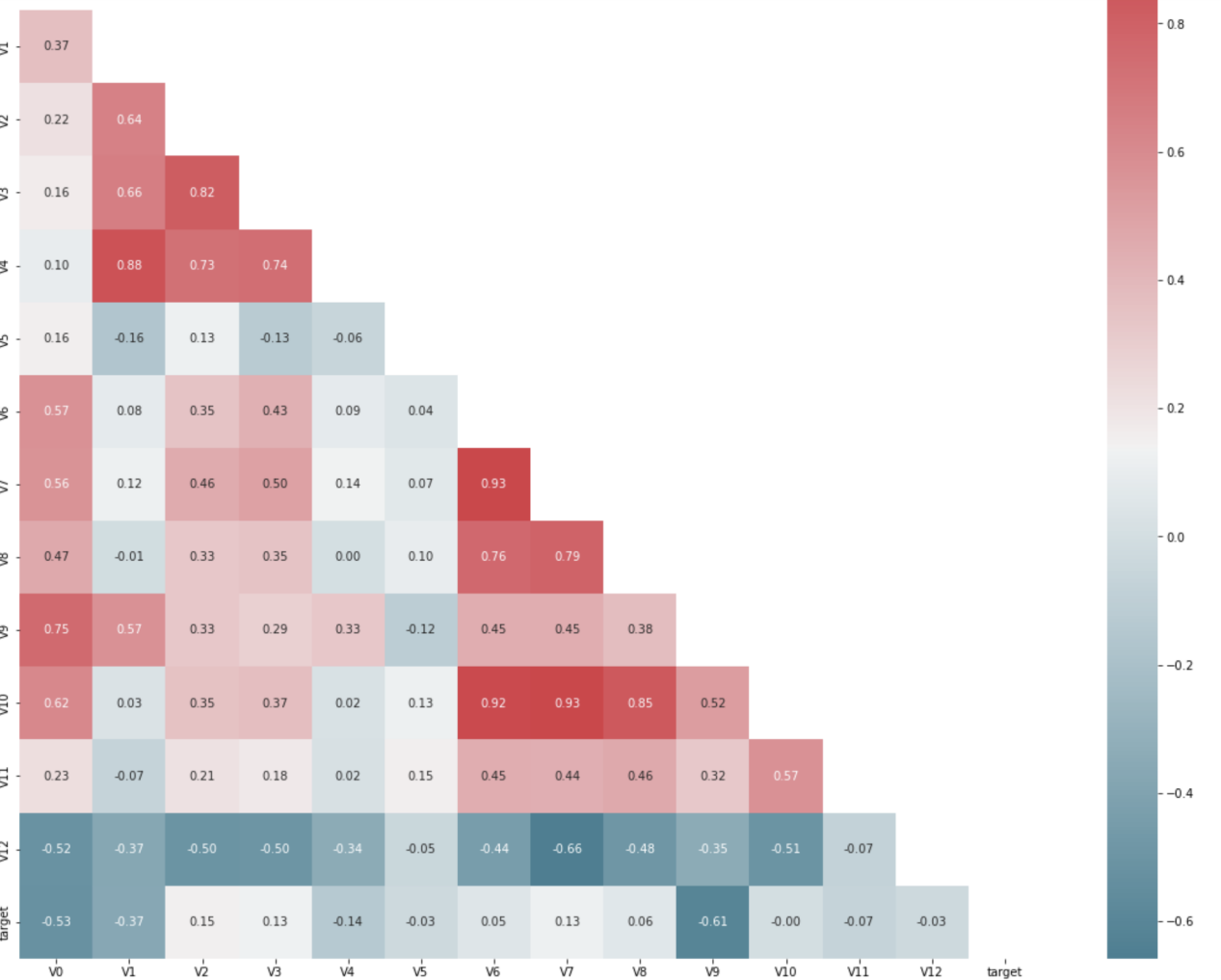


图5-9 梯形热力图

筛选出相关度大于0.1的特征，即会有相关性的特征，即使呈现弱相关。如图5-10所示。

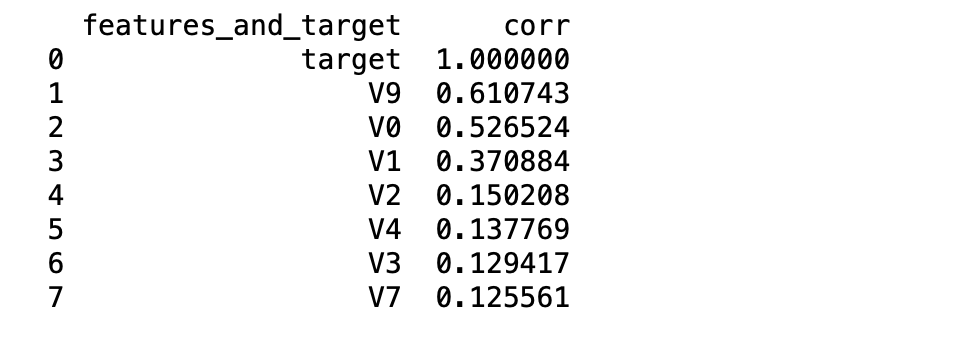


图5-10 相关性呈现

接下来划分数据集，分为训练集和测试集。

首先过滤异常值，此处过滤异常样本的基本思想为：利用岭回归先对train 数据进行建模，建模后的获取 x\_train的预测值y\_train\_predict，若真实值与预测值相减的绝对值 |y\_train-y\_train\_predict| >α\*y\_train.std()，则判定该样本为异常值；此处α参数经过多次测试，取了0.8 较为合适。

将其过程可视化如下：

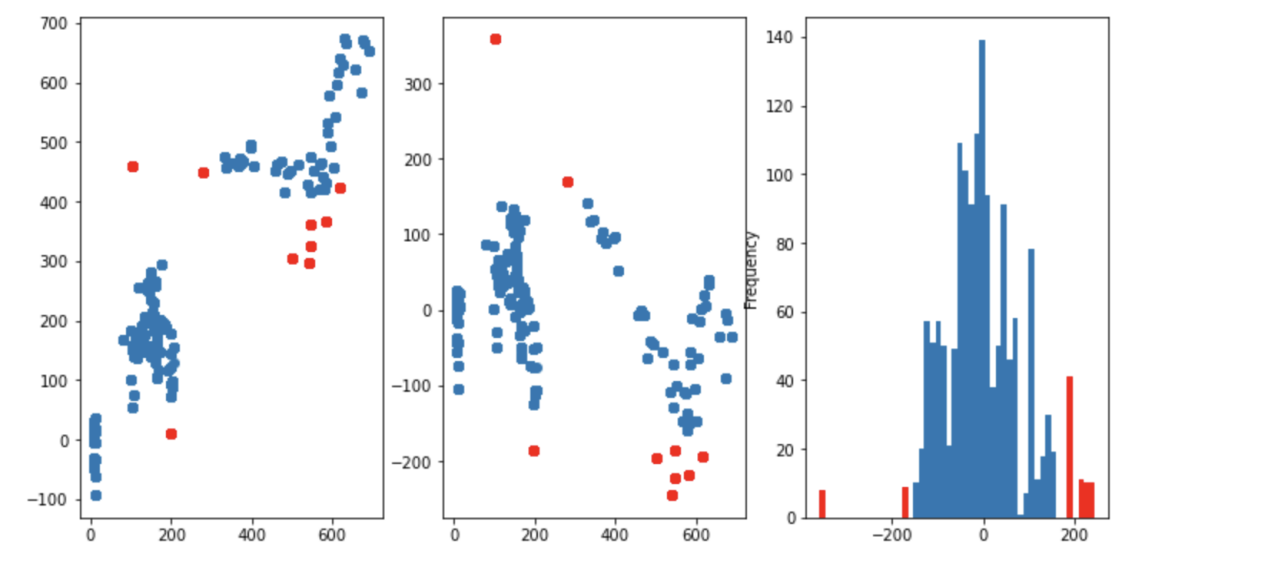


图5-11 岭回归预测

获取行索引，异常值过滤，如图5-12和5-13所示：

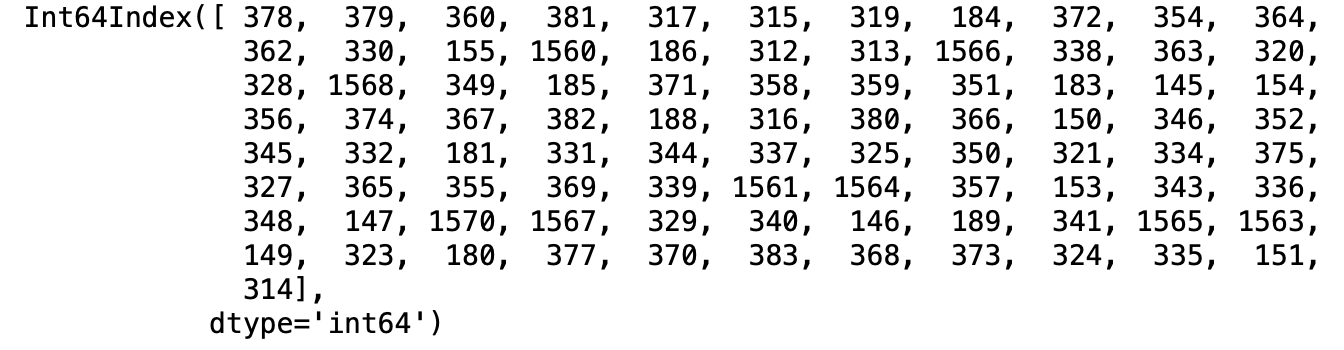


图5-12 异常值索引（1）

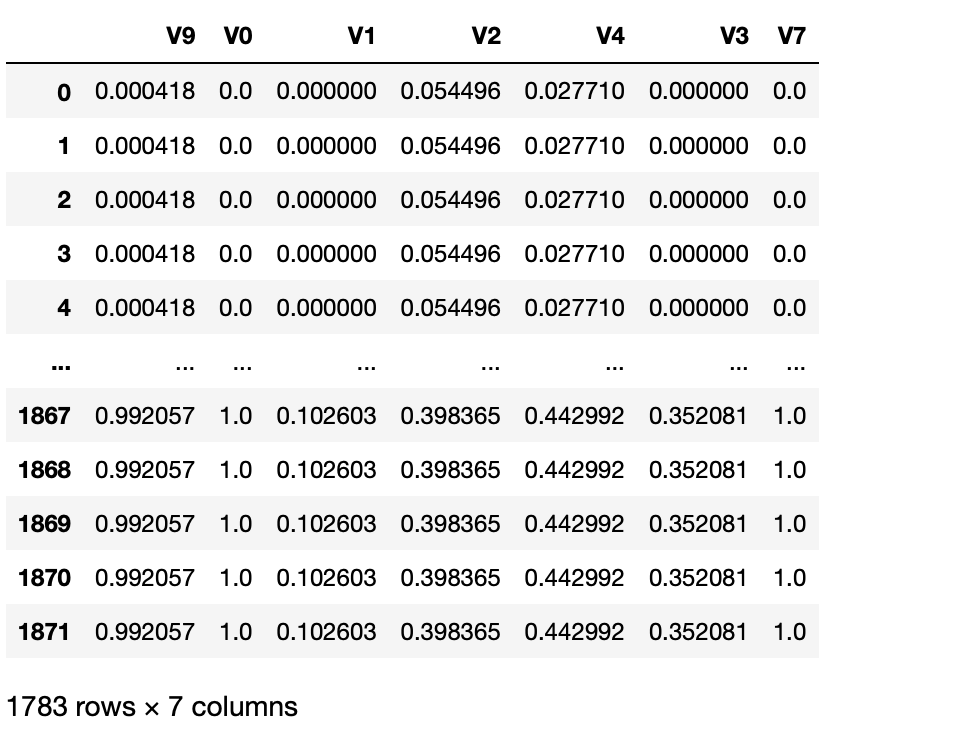


图5-12 异常值索引（2）

## 5.分析与结论

### 5.1信度分析

**1.水稻产量预测**

针对水稻的各地区的年总作物产量，我们使用spss软件进行了回归分析。以下对实验结果进行分析，对实验的结果进行解读。

**由结果得：**本研究纳入的自变量中，柴油使用量，冷冻灾成灾面积，近地面短波辐射，农用氮肥施用折纯量^2，农用氮肥施用折纯量^2，播种面积，温度，近地面风速，近地面空气湿度，近地面气压，近地面长波辐射，降水率，旱灾成灾面积，风雹灾受灾面积，旱灾受灾面积，水灾受灾面积，有效灌溉面积，农用氮肥施用折纯量，农用化肥施用折纯量的P值具有统计学意义，即这些因素是影响作物产量的主要因素。

理论模型：

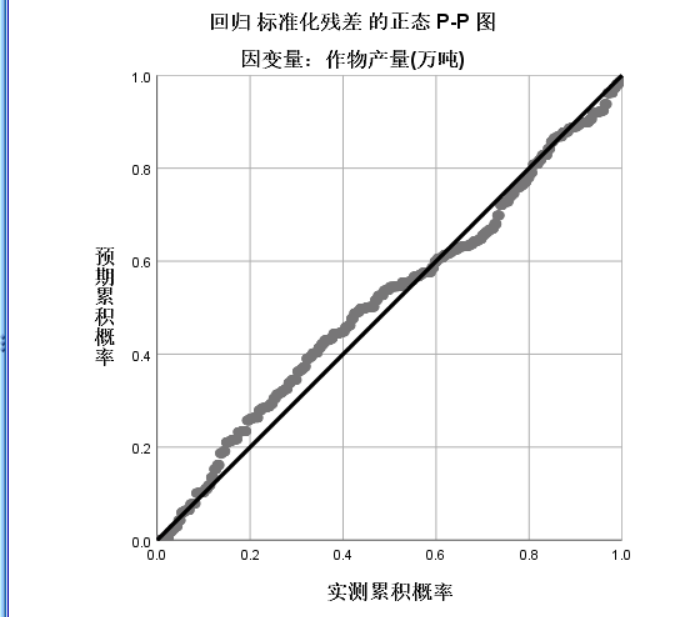
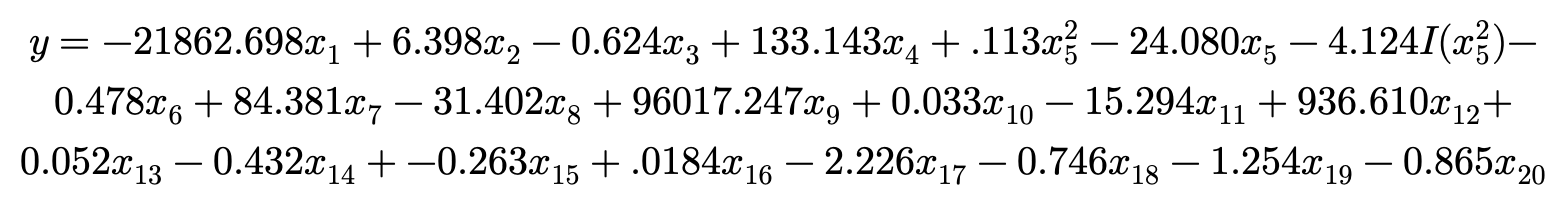


图5-13 回归标准化残差的正态P-P图（水稻）

由回归标准化残差的正态P-P可以看出：残差效果较好，所有的点大致都在一条直线上，可以认为残差符合正态分布的要求。

表5-2 ANOVA表（水稻）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 平方和 | 自由度 | 均方 | F | 显著性 |
| 回归 | 125706653.6 | 22 | 5713938.801 | 477.885 | .000b |
| 残差 | 22107961.98 | 1849 | 11956.713 |  |  |
| 总计 | 147814615.6 | 1871 |  |  |  |

a.因变量：作物产量（万吨）

b.预测变量：（常量），农用化肥施用折纯量（万吨），近地面风速（米每秒），冷冻灾成灾面积（千公顷），风雹灾受灾面积（千公顷），水灾受灾面积（千公顷），近地面短波辐射（瓦每平方米）,降水率（毫米每小时），温度/K,冷冻受灾面积（千公顷），有效灌溉面积（千公顷），风雹灾成灾面积（千公顷），旱灾受灾面积（千公顷），农用化肥施用量折纯量^2（万吨^2），水灾成灾面积（千公顷），近地面气压（帕）,近地面空气湿度（千克^2）,农用氮肥施用折纯量^2（万吨^2），S近地面长波辐射（瓦每平方米），农用氮肥施用折纯量（万吨）

该表格可以用来检验模型的统计学意义，由结果可得F=477.885，P<0.001，因此拒绝原假设，认为回归模型通过了置信水平为0.05的F检验，即所拟合的方程具有统计学意义。

表5-3 变量解释程度与独立性检验（水稻）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型摘要b | | | | | |
| 模型 | R | R方 | 调整后R方 | 标准估算的错误 | 德宾-沃森 |
| 1 | .922a | .850 | .849 | 109.3467549 | 2.072 |

a.预测变量：（常量），农用化肥施用折纯量（万吨），近地面风速（米每秒），冷冻灾成灾面积（千公顷），风雹灾受灾面积（千公顷），水灾受灾面积（千公顷），近地面短波辐射（瓦每平方米）,降水率（毫米每小时），温度/K,冷冻受灾面积（千公顷），有效灌溉面积（千公顷），风雹灾成灾面积（千公顷），旱灾受灾面积（千公顷），农用化肥施用量折纯量^2（万吨^2），水灾成灾面积（千公顷），近地面气压（帕）,近地面空气湿度（千克^2）,农用氮肥施用折纯量^2（万吨^2），S近地面长波辐射（瓦每平方米），农用氮肥施用折纯量（万吨）

b.因变量：作物产量（万吨）

由表格可以看出，调整后的R方为0.849，即总体变量对因变量的解释程度达到84.9%。拟合效果较好，说明模型比较稳定；本研究的得宾-沃森检验值为2.072，一般来说，得宾-沃森检验值分布在0—4之间，越接近于2检验值互相独立的可能性越大。可以认为本研究中的观测值具有相互独立性。

**2.小麦产量预测**

针对小麦的各地区的年总作物产量，我们使用spss软件进行了回归分析。以下对实验结果进行分析，对实验的结果进行解读。

**由结果得：**本研究纳入的自变量中，播种面积，冷冻灾受灾面积，农用氮肥施用折纯量^2，农用化肥施用折纯量^2，降水率，近地面短波辐射，近地面风速，近地面空气湿度，近地面长波辐射，近地面气压，温度，冷冻灾成灾面积，旱灾成灾面积，水灾受灾面积，旱灾受灾面积，有效灌溉面积，农用化肥施用折纯量，柴油使用量，农用氮肥施用折纯量的P值具有统计学意义即：

播种面积，冷冻灾受灾面积，农用氮肥施用折纯量^2，农用化肥施用折纯量^2，近地面短波辐射，近地面风速，近地面空气湿度，近地面长波辐射，近地面气压，温度，冷冻灾成灾面积，旱灾成灾面积，水灾受灾面积，旱灾受灾面积，有效灌溉面积，农用化肥施用折纯量，柴油使用量，农用氮肥施用折纯量是影响作物产量的主要因素理论模型：

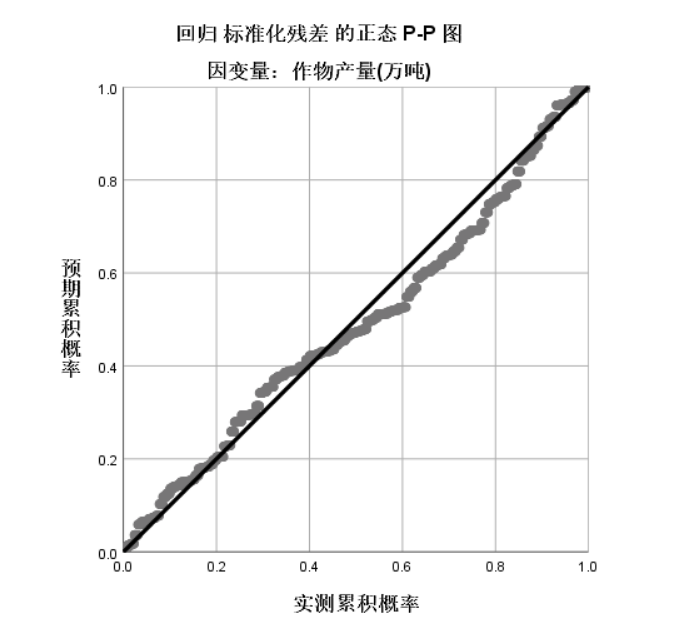
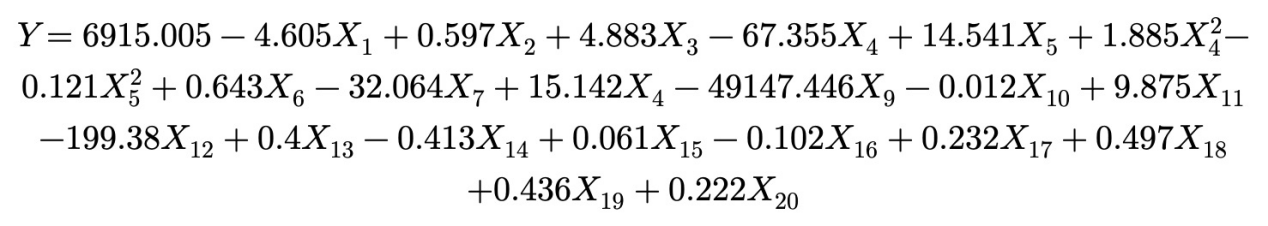


图5-14 回归标准化残差的正态P-P图（小麦）

由回归标准化残差的正态P-P可以看出：残差效果较好，所有的点大致都在一条直线上，可以认为残差符合正态分布的要求。

表5-4 ANOVA表（小麦）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | 平方和 | 自由度 | 均方 | F | 显著性 |
| 回归 | 18675750.67 | 22 | 848806.849 | 427.736 | .000b |
| 残差 | 3050049.507 | 1537 | 1984.417 |  |  |
| 总计 | 21723800.18 | 1559 |  |  |  |

a.因变量：作物产量（万吨）

b.预测变量：（常量），农用化肥施用折纯量（万吨），近地面风速（米每秒），冷冻灾成灾面积（千公顷），风雹灾受灾面积（千公顷），水灾受灾面积（千公顷），近地面短波辐射（瓦每平方米）,降水率（毫米每小时），温度/K,冷冻受灾面积（千公顷），有效灌溉面积（千公顷），风雹灾成灾面积（千公顷），旱灾受灾面积（千公顷），农用化肥施用量折纯量^2（万吨^2），水灾成灾面积（千公顷），近地面气压（帕）,近地面空气湿度（千克^2）,农用氮肥施用折纯量^2（万吨^2），S近地面长波辐射（瓦每平方米），农用氮肥施用折纯量（万吨）

该表格可以用来检验模型的统计学意义，由结果可得F=427.736，P<0.001，因此拒绝原假设，认为回归模型通过了置信水平为0.05的F检验，即所拟合的方程具有统计学意义。

**3. 变量解释程度与独立性检验**

表5-5 变量解释程度与独立性检验

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型摘要b | | | | | |
| 模型 | R | R方 | 调整后R方 | 标准估算的错误 | 德宾-沃森 |
| 1 | .927a | .860 | .858 | 44.54679984 | 2.081 |

a.预测变量：（常量），农用化肥施用折纯量（万吨），近地面风速（米每秒），冷冻灾成灾面积（千公顷），风雹灾受灾面积（千公顷），水灾受灾面积（千公顷），近地面短波辐射（瓦每平方米）,降水率（毫米每小时），温度/K,冷冻受灾面积（千公顷），有效灌溉面积（千公顷），风雹灾成灾面积（千公顷），旱灾受灾面积（千公顷），农用化肥施用量折纯量^2（万吨^2），水灾成灾面积（千公顷），近地面气压（帕）,近地面空气湿度（千克^2）,农用氮肥施用折纯量^2（万吨^2），S近地面长波辐射（瓦每平方米），农用氮肥施用折纯量（万吨）

b.因变量：作物产量（万吨）

由表格可以看出，调整后的R方为0.858，即总体变量对因变量的解释程度达到85.8%。拟合效果较好，说明模型比较稳定；本研究的得宾-沃森检验值为2.081，一般来说，得宾-沃森检验值分布在0—4之间，越接近于2检验值互相独立的可能性越大。可以认为本研究中的观测值具有相互独立性。

具体模型数据可查看《附件五：信度分析模型构建》。

### 5.2 实践案例仿真

以武汉市彭盛源家庭农场为例，经小云鹰系统测得土地面积大小为273.3亩，2020年当地年均亩投入为1352元，预测年产量籼稻共996千克每亩，预估净收入三十一万六千余元，违约距离1.823，预期违约率3.416%。

根据精算模型计算，在损失赔付率为90%、赔付触发条件为综合收入下降超10%的条件下，我们给出的参考保费为每亩30.4元，总保费为8308.3元。

经项目组成员实地考察，与农场主沟通确认实际土地面积268.2亩，2020年均亩投入为1324元，实际产量约976千克每亩，净收入三十万七千余元，经验证发现误差在百分之三左右。

## 6.安装使用

依照上述陈述，本作品分为四个部分：Python 后端、SSM后端、前端、Jupyter分析文件。其中Jupyter分析文件是进行数据处理与数据分析的文件，不需要部署，其他三项都需部署。  
 Python后端需在Python3.7及以上环境运行。通过安装最新版Anaconda，并进入到项目目录下，使用pip工具安装requirements.txt中的依赖包。随后，使用Python分别运行Forrest.py以及FCN.py从而启动图像识别以及预测服务。由于使用了深度学习技术进行高精确的识别和预测，对机器的内存、CPU要求较高，因此部署Python服务时务必确保机器中至少有4GB可用内存。  
 SSM后端采用IDEA2021个人版进行开发。安配置好NodeJS和tomcat后，进入项目目录，执行npm install命令安装依赖，随后执行命令node bin/www启动后端服务。  
 前端采用VUE技术打包，需要安装NodeJS。安装后，进入项目目录，执行npm install命令安装依赖，随后执行命令npm start启动前端服务器，使用浏览器访问本地8080端口浏览页面。  
 后端依赖Redis、MySQL等数据库，数据的安装与部署不再赘述。

## 开发体会与作品总结

在开题阶段，项目组成员对农业相关知识缺乏了解，且对国家政策和普惠金融相关概念认识不足。通过翻阅今年的红头文件、一号文件，参考三农专家温铁军老师的建议等等，并查阅大量相关论文，对我国农业国情进行了大量分析，最后得出精确用户画像，确定了以To B业务为主，利用基于计算机视觉与机器学习的遥感农业大数据手段为农村弱信用群体进行征信的核心开发方向。

在数据收集阶段，由于国内许多农业和遥感数据相关网站多用于科研领域，想获得精确数据需要国家级科研项目认证，队员们不得不去国内外广大开源数据网站上搜集可用数据，并且寻找可用的卫星地图api以及可见光和红外光谱遥感卫星影像。最后在指导老师的帮助下队员们搜集到了初期数据，利用spss进行信度分析之后发现许多数据的相关性不足，最后进行多轮数据清洗确定了现有的自变量。

由于组内成员均为计算机学院学生，对遥感数据处理不甚了解，组员们大量查阅相关论文，寻求与农场合作的机会，在否决掉光谱分析等处理手段后确立了ndvi为长势评估的核心手段，并且搜集到了相关农场信息。

数据处理和模型训练的过程中，最开始配环境的时候，用于机器学习的sklearn库一直安装失败，在反复试验了多次之后，发现与别的包版本不兼容，再进行lib包版本调试后最终安装好，在跑预测数据时，因为图像数据量大造成卡顿，等待许久才成功。

由于数据拟合程度不够，我们使用了多种模型交叉验证的方式，最后选取LGB模型训练结果作为我们产量估测值。

本项目旨在为农村弱信用群体征信提供合理解决方式，最终以基于计算机视觉和机器学习的遥感农业大数据新应用的形式呈现。在开发过程中，团队成员齐心协力紧密合作，多次实现跨领域跨学科式的突破自我，汲取相关知识。在巩固和掌握原有计算机技术的基础上，学习了机器学习、web开发、ui设计等新技术，并运用到实践中。在各个学科交叉领域跨域交流的过程中，提升了自己的交流能力，学习到了其他学科思维和解决问题的方法。在团队合作的过程中虽然有过争吵，但更多的是无数思维火花的碰撞，是一起熬夜也要解决问题的友谊。

目前本项目还有许多有待实现的拓展功能：针对商业模式，开辟新的To C方业务；扩充自有数据集，改良算法与模型等等；在商业推广方面运营自媒体，通过线上线下相结合的营销方式实现裂变式吸纳流量。这次项目开发经历令我们受益匪浅，或许比赛只是我们的起点，我们的团队合作、团队精神永远不会改变。我们将践行自强弘毅、求是拓新的精神，将更多的精力投入后续功能开发以及性能提升中，与小云鹰一起不断进步。