一 Introduction
问题背景
问题重述
模型准备
模型假设
二 数据描述性统计
特征变量总结
对“Lab Status”的分析
缺失信息统计
地区分布数据统计
时间分布统计
三 数据预处理
易混淆物种的特征关键词提取
分词并去除停用词
构建语料库
计算TF-IDF值
群众notes与特征关键词的相关度
分词并去停用词
计算TF-IDF值
处理特殊数据
图像相关度
卷积神经网络模型原理介绍
卷积神经网络模型构建
利用卷积神经网络计算相似度
发现日期
发现地点与Vespa mandarinia分布之间的距离
四 蔓延情况预测
五 逻辑回归模型
逻辑回归原理
构建逻辑回归模型
六 逻辑回归模型统计学检验
假设检验
区间估计
拟合优度
灵敏度分析
ROC曲线分析
七 正面目击报告筛选
八 模型更新
九 问题五模型构建
模型优缺点
十 备忘录

一 Introduction

问题背景

Vespa mandarinia是世界上最大的胡峰。他有极强的毒性，在极端情况下可能会致人死亡，同时它具有极强的攻击性，会攻击其他的蜜蜂并在几小时内摧毁一个蜂巢。他原本生长于亚洲大陆，又称Asian giant hornet，在其他大陆，比如美洲，因为没有天敌的存在他能够快速的繁衍蔓延并破坏当地的蜂群系统，对当时的生态系统造成费仲严重的破坏，所以对于有Vespa mandarinia入侵的国家，如何有效的消灭所有的Vespa mandarinia是一个非常重要任务。

2019年10月，工人加拿大的温哥华岛发现了一个Vespa mandarinia的巢穴，这个巢穴很快就被摧毁但是从这以后在与温哥华岛相邻进的美国华盛顿州又陆续发现了几只Vespa mandarinia。这个事实证明Vespa mandarinia已经入侵美国华盛顿州，为了保护当地的生态以及人们群众的生命安全，美国政府必须采取行动，利用有限的资源来消灭Vespa mandarinia。

为了充分利用资源，美国政府开通了群众报告热线并建立了一个网站，群众可以给该热线或者网站发送报告用以汇报其在日常生活中发现的Vespa mandarinia。但是因为群众并不专业，他们很多时候都错误的把其他昆虫认为是Vespa mandarinia，这就需要专业人士对这些报告进行筛选，找出确定是Vespa mandarinia的报告或者无法完全否定是Vespa mandarinia的报告。之后派遣专业人士前往这些报告中记录的观察到Vespa mandarinia的地点进行进一步的搜寻与消灭工作。

问题重述

the Washington State Department of Agriculture为了从群众的报告中挖掘出更有效的信息，提高报告数据的价值密度，希望我们对群众的报告进行进一步的处理与分析。我们的数据挖掘过程应该包含一下几个方面：

1. 预测大黄蜂在一段时间内的蔓延情况并判断预测精度。
2. 利用所提供的数据和图像文件，建立预测其他大黄蜂被误认为是亚洲巨峰的可能性的错误分类模型。
3. 使用所建立的模型，筛选最有可能的正面目击报告进行调查。
4. 随着时间推移，新的报告加入，解决模型的更新问题，以及多久更新一次。
5. 使用所构建的模型，分析构成华盛顿消灭害虫的证据。

在完成这些要求后，the Washington State Department of Agriculture希望我们能够做一个memorandum用于向他们汇报我们的结果。

模型准备

模型假设

为了简化我们的模型，我们做了一下一些主要的假设：

1.negative的样本都是Vespa mandarinia的捕食对象。

2.每一次报告的人员对昆虫的体态特征（大小、头部和身体颜色等）的估计都是准确的 。

3.报告提交的时间分布均匀。

4.确认某一地区存在Vespa mandarinia后相关人员会前往捕杀当地捕杀Vespa mandarinia，捕杀后可能有极少数的Vespa mandarinia逃脱捕杀，但蜂后不可能逃脱。

二 数据描述性统计

相关部门为我们提供的了用户报告数据，包括一份记录群众报告基本信息的Excel表格数据，每份报告附带的图像数据以及一份记录了图像数据与报告对应关系的Excel表格数据。同时还未我们提供了一篇Vespa mandarinia的介绍文档，介绍了Vespa mandarinia的习性和一些易与Vespa mandarinia混淆的物种。接下来我们的论文都是基于相关部门为我们提供的这些数据。在本节中我们将对群众报告基本信息数据集合图像数据与报告对应关系数据集进行一个简单的描述性统计用于初步探究所提供的数据。

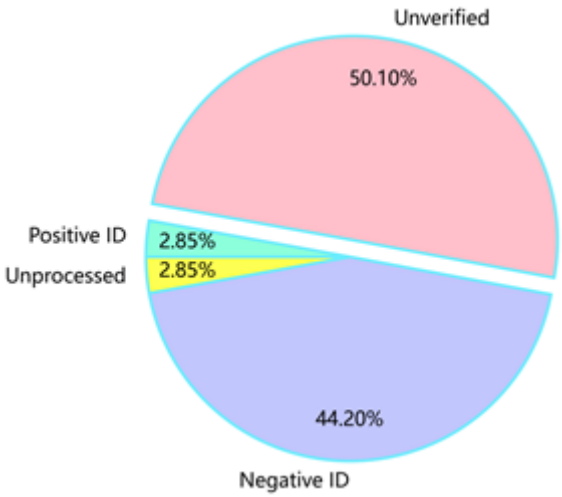
特征变量总结

群众报告基本信息数据集包含8个特征变量，分别为:" GlobalID"," Detection Date"," Notes"," Lab Status"," Lab Comments"," Submission Date"," Latitude"," Longitude"。其中" GlobalID"是每一份报告数据独有的标志；"Detection Date"是群众发现那种可能是Vespa mandarinia的物种的时间；"Notes"是群众对自己发现的物种或者提交的报告的一个描述；"Lab Status"是政府机构给这份报告的一个标记，用于表示报告中的物种是否是Vespa mandarinia。如果确实是Vespa mandarinia就会被标记为"Positive ID"；如果被证实不是Vespa mandarinia就被标记为"Negative ID"；如果不能确定是否是Vespa mandarinia则被标记为"Unverified"；如果报告正在处理中则被标记为"Unprocessed"。" Lab Comments"是政府机构对群众上传的报告的一个评价，可能会包括政府对上传报告的群众的感谢和群众错认的原因等内容。" Submission Date"是群众提交报告的时间；" Latitude"和" Longitude"分别指群众发现那种可能是Vespa mandarinia的物种的经度与纬度。

图像数据与报告对应关系表包括3个特征变量，分别是:"FileName"," GlobalID"," FileType"。其中"FileName"是每个图像文件的名称。" GlobalID"是每个图像文件对应的报告ID。"FileType"是每个图像文件的类型，包括图片，压缩包和视频。

对"Lab Status"的分析

我们已经知道"Lab Status"有四个值现在分别统计着四个值出现的次数在总次数中的占比，统计结果如下图所示

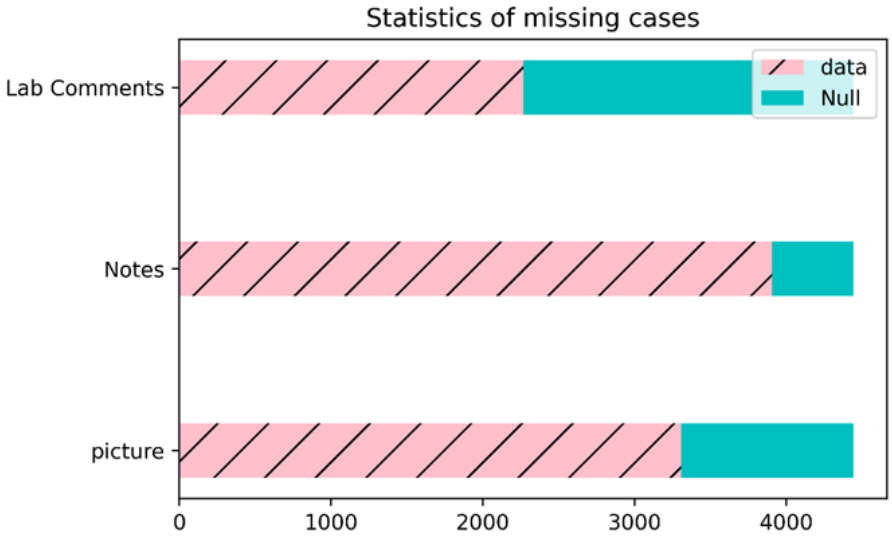


从上图可以看出值为"Positive ID"的报告有14份，值为"Negative ID"的报告有2069份，值为"Unverified"的报告有2342份，值为"Unprocessed"的报告有15分。

从上图中可以值为"Positive ID"的报告仅占总报告的2.85%，值为"Negative ID"的报告占总报告的44.20%，值为"Unverified"的报告在总报告中的占比高达44.20%。从这个结果中我们可以发现群众提交的报告有一半因为信息确实无法有效识别还有接近一半的报告已经被证实是错误的将别的物种认为是Vespa mandarinia。这是因为相对于政府部门的专业人士，普通群众缺乏相关知识，所以无法正确识别Vespa mandarinia也无法为专业人士提供准确、清楚的有用信息。所以仅有极少部分的报告正确的为相关部门提供Vespa mandarinia的信息，帮助相关部分消灭Vespa mandarinia。但是因为大量无用报告（"Unverified"和"Negative ID"）的存在，政府必须浪费大浪的资源来处理这些报告，所以这就需要我们提供一种能够帮助政府部分快速有效的是别处正确的报告的模型来提高资源利用率。

缺失信息统计

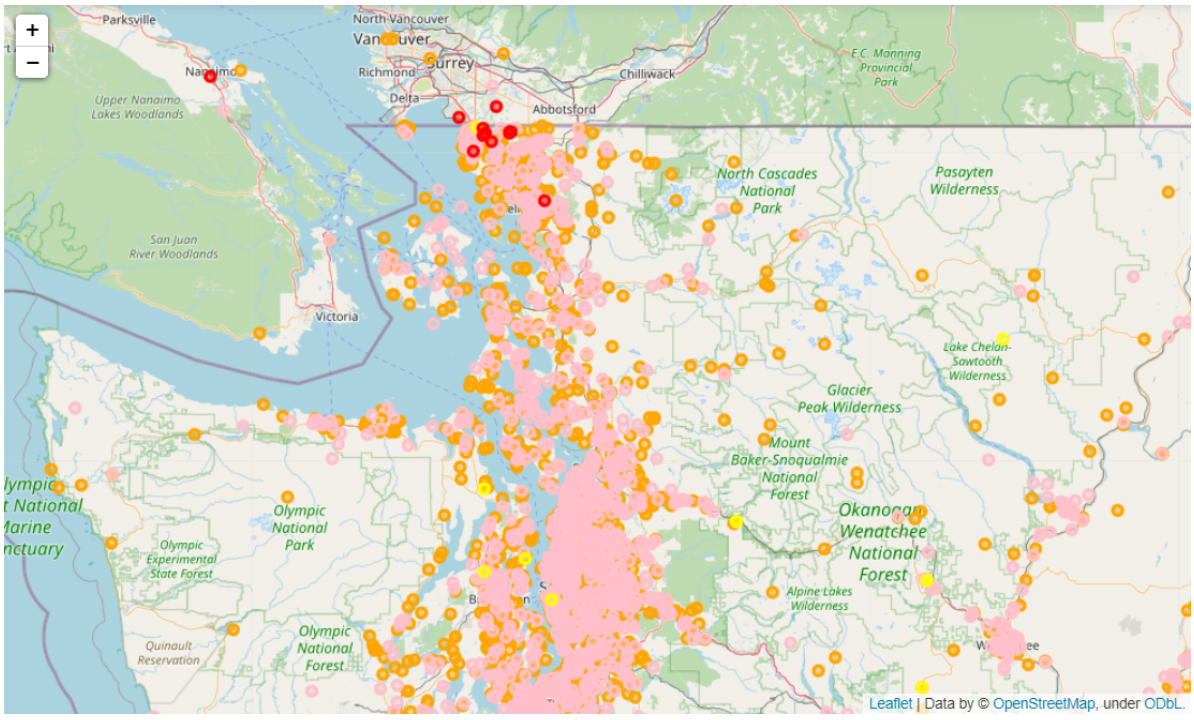
相关部门无法判断是否是Vespa mandarinia的很大一个原因就是信息的缺失，这里主要是"Notes"信息和附带的图像文件信息，这两个信息对我们后续的模型构建有非常重要的作用，所以在这里我们对其的数据缺失情况进行一个统计。同时"Lab Comments"也是后续模型需要的一个重要因素，我们将其放在一起展示，统计结果如下所示：



上述条形图描述了picture，Notes和Lab Comments三类有效数据和空缺值得数量，其中picture有效数据3306条，空缺值1135条，Notes有效数据3904条，空缺值537条，Lab Comments有效数据2266条，空缺值2175条。

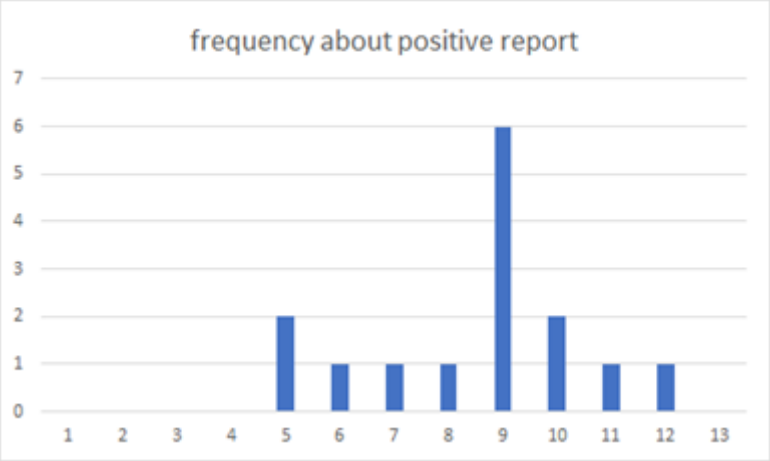
地区分布数据统计

如下图所示，我们将每份报告地区的分布绘制到地图上。我们发现被确认存在Vespa mandarinia的地区较为集中，大致分布在经度48到49，纬度-124到-122的区域内。其他提交的报告，包括错认的报告和无法判断的报告，则分布得较为分散，但是绝大部分也是位于水域边缘，这是因为根据虫类的习性，相较于内陆地区水域周围更容易生长虫类。



时间分布统计

根据文献记载，绝大部分Vespa mandarinia会在冬天逐渐死亡蜂群中的皇后则会进入土壤中休眠，直到第二年的春天受精的皇后才会重新开始活动并重新生育其他的Vespa mandarinia。文献资料表示一个蜂群中的Vespa mandarinia的数量还在8月达到巅峰。这些文献资料充分的说明了Vespa mandarinia的数量与季节或者月份有关。如果某一个月份Vespa mandarinia的数量越多那么，那么在那个月份群众就越有可能发现Vespa mandarinia，所以我们可以得出一下结论：夏季和秋季（或者说7月份之后）更容易观测到Vespa mandarinia。为了验证我们的推测我们对被确认是Vespa mandarinia的14分报告的发现时间按照月份做了统计，统计结果如下所示：



从上图中我们可以发现前四个月都没有发现Vespa mandarinia，发现最多Vespa mandarinia的是9月份，大部分的Vespa mandarinia都在夏季和秋季（6月22日至12月21日）被发现，这与我们的推测相呼应。

三 数据预处理

在本题中我们最终需要得出的是每一分报告是错认的概率并根据设定的阈值将报告划分为“Negative ID”或“Positive ID”，所以我们并不关心 “Unverified” 与 “Unprocessed”的报告。基于此，我们分别提取出整个群众报告基本信息表中被标记为“Negative ID”与“Positive ID”的数据。其中被标记为“Positive ID”的数据仅有14条，但是被标记为“Negative ID”的数据却有2069条，这样的数据显然存在分布不平衡问题，对我们最终的模型会存在影响。

根据上述的筛选过程，我们新获得两份数据集，其中一份数据集仅包含被标记为“Positive ID “的14条数据，另一份数据集包含被标记为” Negative ID “的2069条数据。这两份数据集都包含” GlobalID “、” Detection Date “、”Notes “、”Lab Status “、”Lab Comments “、”Submission Date “、”Latitude “、” Longitude “这八个特征。我们后续的数据预处理过程仅处理这2083条数据。

易混淆物种的特征关键词提取

浏览数据后我们发现在多个被标记为“Negative ID“的报告数据拥有” Lab Comments “这个特征值。Lab Comments是专业人士对群众提交上的报告的一个评论，在这个评论中我们发现专业人士不仅对发送报告的群众表示感谢还给出了被错认为是Vespa mandarinia的物种与Vespa mandarinia的差别或者被错认的原因，我们用5条数据来进一步阐述这个发现，5条数据如下表所示：

众所周知，群众错认的原因一般来说就是两种物种在某些方面过于相似，比如因为大小、形状、颜色的相似，Vespa crabro就经常被误认为是Vespa mandarinia。同样，如果不做特别的区分，我们的模型也可能会错认。所以我们需要一个参照集，如果群众报告中的Note或者图像文件包含参照集中的特征那么这个报告有更高的概率是错认的。幸运的是，我们通过观察数据已经发现”Lab Comments “中包含专业人士给出的易混淆物种区别于Vespa mandarinia的特征，所以通过提取这些特征的关键字，我们可以构建一个易混淆物种的特征关键词表，用于后续的数据处理与模型建立过程，该表详细的用处可以参考后两小节的内容。

构建易混淆物种关键词表的过程如下所示：

分词并去除停用词

首先，我们将包含所有被标记为“Negative ID“的数据集中所有的” Lab Comments “以每一条数据为单位提取出来，然后将每一个句子按照词与词之间的空格划分开，得到一个以词为单位的数据集，在这个数据集中，属于同一个” Lab Comments “的词会被存放在同一个子集中。之后我们考虑到英文书写过程中的大小写问题，我们将所有的词都转化为小写形式存储。

之后，我们知道在一个句子中总会有一些没有什么实际含义或者或者实际含义很少的词汇，比如感叹词，这些词对我们或许的处理并没有什么显著性的帮助，有时还会干扰我们的结果。这样的词就称为停用词，在我们的处理过程中我们选择将这些词删除。我们删除停用词的的方式是从Python的第三方库nltk中下载一个包含1286个停用词的停用词表，然后将数据集中的每一个子集中的每一个词与停用此表中的词相比较，如果数据集中的词也同样存在停用此表中那么就将这个词删除，否则继续保留这个词。由此我们得到一份新的数据集，这个数据集是原先数据集的子集。

构建语料库

为了计算后续每一个句子中每一个词的TF-IDF值，我们需要构建一个语料库，这个语料库应该包含处理后的数据集中所有的词。同样，这个语料库中属于同一个句子的词需要放在同一个子集中。由于去停用词后获得的数据集完全符合要求，这里我们并不需要做其他操作，只需将去停用词后获得的数据集直接当做语料库即可。

计算TF-IDF值

计算每一个句子中每一个词的TF-IDF值。

TF-IDF算法的主要思想是：如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，能够作为这一篇文本的关键词。其中TF是指词频即一个词在一个文本数据中出现的频率，IDF是逆文档频率指数是一个词普遍重要性的度量。某一个词的IDF可以由总文本数据数目除以包含该词的文本的数目，再将得到的商取以10为底的对数得到。将某一个词的TF值与IDF值相乘就可以得到某一个词对应的TF-IDF。显然，一个词的TF-IDF值越大，这个词就越重要。具体计算公式如下所示：

以上式子中 n*{i,j} 是该词在文件*d{j}中的出现次数，而分母则是在文件d_{j}中所有字词的出现次数之和。|D|：语料库中的文件总数

$|\{j : t_i \in d_j\}|$：包含词语t_i的文件数目（即$n_{i,j} \neq 0$的文件数目）如果该词语不在语料库中，就会出现被除数为零的情况，因此我们使用$1 + |\{j : t_i \in d_j\}|$

然后可以得到

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i}$$

在了解TF-IDF算法基本原理的基础上，我们利用起数学公式计算第三步所得到的语料库中的所有词的TF-IDF值。

首先，我们对于每一个句子，即语料库中的每一个子集，统计子集中每一个词出现的频率；之后统计出每一个词在该语料库中的多少个子集中出现以及总的子集数。得到这三个数据后就可以按照上述公式（3）计算每一个词的TF-IDF值。注意，因为每一个句子中即子集中可能存在相同的词，所以同一个词会有多个TF-IDF值，并且在不同的句子中该词的TF-IDF值不相同，对于这一类词，我们使用所有TF-IDF

的平均值作为他们的TF-IDF值。

经过上述操作我们已经获得一个包含786个词以及每个词对应TF-IDF值得数据集。浏览发现，有一些词并不能很好的描述易混淆物种与Vespa mandarinia之间的差别，不应该出现在我们的特征关键词表中。同时我们也发现这一类词的TF-IDF值普遍小于0.5，所以我们以0.5为阈值，将获得的数据集中所有TF-IDF值小于0.5的词删除，最终留下298个词。另外我们考虑到英文书写的过程中动词有六种形式（动词原型、第三人称单数、过去式、现在进行时、过去分词），名词有两种形态（原型、复数），所以我们将原来的298个词中出现的动词和名词的其他形态也添加到该数据集中，这些新添加的词的TF-IDF值与原来的数据集中对应的那个词的TF-IDF的值相同。最终完美得到一个包含352个词的特征关键词表，部分关键词及其TF-IDF值如下所示（具体的特征关键词表见附录）：

群众notes与特征关键词的相关度

在上一小节中我们得到了一个易混淆物种特征的关键词表，表中包含352个词和每个词对应的TF-IDF值。浏览数据，我们发现绝大部分的报告都会有一个notes，这个notes是群众对自己所发送的这份报告的文字描述，有些notes会包含群众对他发现的物种的特征。显然当某一份报告的notes与我们的易混淆物种特征关键词表有较高的相似度时，这份报告是错认的概率就会相对提高。所以我们通过计算每一篇报告与易混淆物种特征关键词表的相关性得出一个新的特征变量，我们讲这个变量称为“Notes_similarity”。

我们需要查询每一个Notes与易混淆物种特征关键词的相似度，这是一种典型的文本相似度度量，最常用的方法就是结合TF-IDF值与余弦相似度。

余弦相似度认为可以使用两个向量之间的夹角的余弦值来度量这两个向量的相关性，根据余弦的特性（0度角时为1,180度角时为-1），显然余弦值越大两个向量就越相似。在实际操作中，我们需要将文本数据通过某种特定的方法映射到两个多维向量中，之后利用如下所示余弦公式得到余弦值:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}.$$

在该公式中，A和B是两个向量，Ai和Bi分别是A和B的各个分量

在我们的实际操作过程中我们需要计算每一个Notes与易混淆物种特征关键词表的相似度，因为在上一小节中我们已经得到易混淆物种特征关键词和每个词对应的TF-IDF中，这个文本数据已经被映射到向量空间中，所以我们同样利用TF-IDF值将Notes映射到向量空间中。我们的步骤如下所示：

分词并去停用词

首先，我们将每一个Notes进行分词并将每一个词都转换成小写形式与去停用词的操作，这个步骤与4.1.1小节一致。

计算TF-IDF值

第二，计算每一个Notes中每一个词的Tf-idf值，计算TF-IDF值得步骤也与4.1.1小节一致。但是，为了方便计算余弦值，我们需要使TF-IDF值映射出的向量的维数都相等。所以我们构建一个新的语料库，这个语料库包含所有的易混淆物种特征关键词和将所有Notes分词并去停用词后获得的所有词，注意在这个语料库中每个词只出现一次，如果出现词重复就仅保留一个词。在我们的时间过程中，这个新的语料库包含429个词。

有了这个语料库之后我们将每一个Notes分词并去停用词后的结果与该语料库进行对比，一旦发现某一个语料库中的词并没有出现在Notes的处理结果中就将这个词添加进去。这里需要注意的是每一个向量中词的分布顺序必须与语料库中词的分布顺序相同，这样的处理是为了使所有向量的每一个分量都有固定的相同的含义。因为我们添加进去的词实际上并没有出现在Notes中，我们令它的词频为0，那么最终这些词的TF-IDF值也为0，对最终的余弦值没有影响。同样，我们也检查易混淆物种特征关键词表，添加缺少的词并使得词的顺序与语料库中词的顺序一致。

最终这一步完成后我们得到14+2067+1=2082个429维的向量，即14个被标记为“Positive ID”的数据的“Notes”映射出的429维向量加上2067个被标记为“Negative ID”的数据的“Notes”映射出的429维向量再加上1个易混淆物种特征关键词映射出的429维向量。

处理特殊数据

在描述性统计中我们发现并不是所有的报告都有一个Notes，对于那些没有提交Notes的数据，按照第二步的计算过程我们得到的“相似度”均为0，这显然不符合实际。所以我们将和他一起被标记为同一类的所有数据的“相似度”的平均值作为他的值，比如当一个被标记为“Positive ID”的数据不存在Notes的值时，就将该数据集中所有被标记为“Positive ID”并且包含“Notes”的数据的“相似度”的平均值作为这个数据的“相似度”。

最终完美得到所有数据的“相似度”，部分数据的“相似度”如下表所示：

图像相关度

图像数据是群众在报告中附带的有关那个可能是Vespa mandarinia的物种的视频或者图片，相较于“Notes”，图像数据更能够直观的展示该物种的特征。根据已经给出的数据，我们可以将图像数据也分成两类，一类是被标记为“Positive ID”的图像，一类是被标记为“Negative ID”的图像。利用卷积神经网络我们可以进行图像识别，分别识别出被被标记为“Positive ID”的图像中的物种特征以及被标记为“Negative ID”的图像中的物种特征。有了这个模型之后，每当群众提交一份新的图像数据我们就可以计算这个图像与被标记为“Negative ID”的图像的物种特征之间的相似度。显然，相似度越高，这份报告就越有可能是错分的。因此，我们可以得到一个新的，可用于判断报告是否错分的特征，我们将其命名为“图像相似度”。

为了计算“Image_similarity”，我们使用卷积神经网络模型进行图像识别。

卷积神经网络模型原理介绍

卷及神经网络模型是一种常见的神经网络模型，它具有局部连接、权值共享的结构特点。局部连接即浅层网络中仅有部分神经元与深层的神经元连接，权值共享即同一层的所有神经元共享相同的连接权重。由于存在这两个结构特点，卷积神经网络的参数规模大大降低，并且能够用于挖掘图象中的局部特性，所以我们使用卷积神经网络模型计算相似度

卷积神经网络的关键是卷积运算，他的基本运算过程是，利用一个被称作卷积核的矩阵，和输入图像矩阵的对应位置的子矩阵作element-wise乘积运算，得到一个常数灰度值，卷积核在输入图像矩阵上每移动一个步长，就会运算一次得到一个相应的灰度值，当卷积核遍历完图象矩阵长和宽方向的所有子矩阵后，会得到一个由一系列灰度值组成的矩阵，这个矩阵就是卷积运算的输出，通常情况下，这个矩阵具有明显的物体边缘特征。

常常在卷积神经网络的基础上组合出更复杂的网络架构，来提高准确率，我们针对大黄蜂的数据集采用了ResNeXt网络架构。

卷积神经网络模型构建

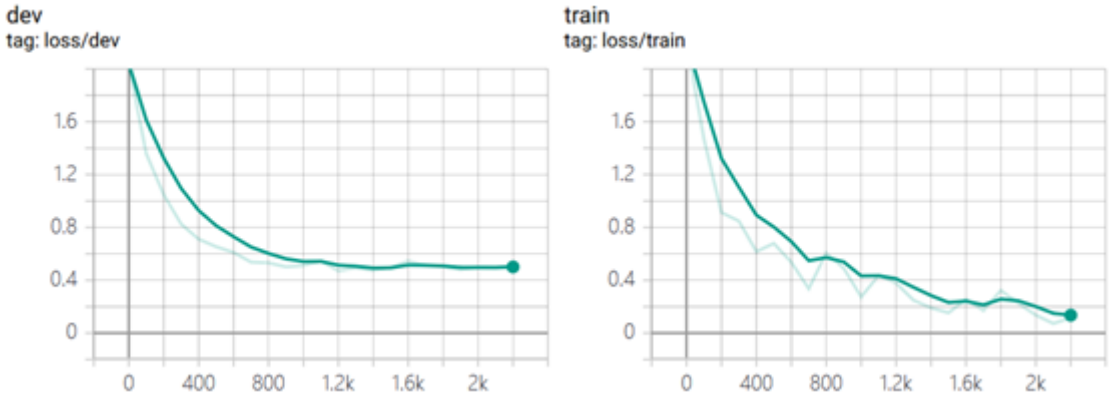
可以画个算法框

因为群众提交的图像数据中包含视频与图片，所以我们截取每一个视频数据中较为清晰的一幕，使得图像文件中仅包含图片数据。之后我们将所有的调节所有图片的长宽比，使其为3:1。最后按照报告的标记，将图像分成两类即Positive 类和Negative 类，并输入到卷及神经网络模型中，模型的输入输出如下所示：

输入：图像

输出：Positive或者Negative

之后我们将70%的数据划分为训练集，30%的数据划分为测试集，再进行模型训练，并通过调节参数使得模型有一个较好的拟合度，即有一个较小的损失率。拟合过程中，模型在训练集和测试集上的损失率变化如下图所示



最终我们的模型在测试集上能够达到0.638的损失率。从图中我们可以看出模型的损失率在拟合的过程中被降低了很多，但是最终的模型在验证集上依然有很大的损失，此时我们模型的参数如下表所示：

stage	output	ResNet-50	**ResNeXt-50 (32×4d)**
conv1	112×112	7×7, 64, stride 2	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2	3×3 max pool, stride 2
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times3$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128, C=32 \\ 1\times1, 256 \end{bmatrix} \times3$
conv3	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times4$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256, C=32 \\ 1\times1, 512 \end{bmatrix} \times4$
conv4	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times6$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, C=32 \\ 1\times1, 1024 \end{bmatrix} \times6$
conv5	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times3$	$\begin{bmatrix} 1\times1, 1024 \\ 3\times3, 1024, C=32 \\ 1\times1, 2048 \end{bmatrix} \times3$
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params.		$\mathbf{25.5}\times10^6$	$\mathbf{25.0}\times10^6$
FLOPs		$\mathbf{4.1}\times10^9$	$\mathbf{4.2}\times10^9$

利用卷积神经网络计算相似度

训练好模型后我们可以向卷积神经网络输入图片，卷积神经网络会输出一个二维向量（x,y），其中x表示这张图片被标记为Positive的概率，y表示这张图片被标记为Negative的概率。显然y的值越高，x的值就越小，那么附带这张图片的的报告就越有可能是错认的。虽然通过y的值我们可以直接得出收否错认，但是因为我们的模型损失率较高，直接使用的话会存在较大的误差，所以我们仅选择将模型计算出来的某一张图片被标记为Negative的概率作为一个判断报告是否错认的特征。我们将这个特征命名为“Image_similarity”。

发现日期

再描述性统计中我们已经发现Vespa mandarinia在6-12月份活动的更加频繁，所以，我们认为群众上交报告中记录的发现可能是Vespa mandarinia的某一物种的时间也可以作为判断这份报告是否错认的一个依据。我们提取出每一分报告中的“detection date ”中的月份并计算在这个月份发现真正的Vespa mandarinia的概率。我们将这个概率作为一个新的特征，命名为“time_probability”。

我们将“detection date ”特证进行处理，仅保留月份信息，得到一列新的数据“month”，并根据图计算出每个月份的概率。对于已经发现Vespa mandarinia的月份，我们将该月份发现的Vespa mandarinia的数量占所有Vespa mandarinia的数量的比例作为这个月的概率。对于暂未发现Vespa mandarinia的月份，即1，2，3，4月份，所以我们采取按季度取均值的方法，给1，2，3，4月份赋予概率，1，2月同12月的概率，4月份同5月的概率。至此所有月份均有一个发现正真的Vespa mandarinia的概率值，如下表所示。（做个表吧，感觉表格清楚一点）

将概率值与所在月份一一对应得到“time_probability”这个特征变量。

发现地点与Vespa mandarinia分布之间的距离

我们假设每确认一只Vespa mandarinia政府相关部门就会前往该Vespa mandarinia所在地进行除虫工作，这个工作不仅仅是除掉那一只Vespa mandarinia，而是希望能够摧毁整个蜂群，因为Vespa mandarinia是群居的。但是考虑到Vespa mandarinia是活动的，根据文献资料显示，每一只工蜂可以在不超过8公里的范围内进行捕食等活动。所以我们认为在当地蜂群中有几只Vespa mandarinia可能会侥幸逃脱捕杀，也就是说被观测到确认存在Vespa mandarinia的那个地方在未来一段时间内仍然有可能存在Vespa mandarinia。另外，文献资料也显示Vespa mandarinia是一种一年一度的物种，每一个Vespa mandarinia群落里有一个皇后，只有受精的皇后才能生育其他的Vespa mandarinia。每一年冬天除了皇后以外的Vespa mandarinia都会死亡。所以为了简化计算，我们假设侥幸逃脱的Vespa mandarinia并不是蜂群中的皇后，那么那些侥幸逃脱的Vespa mandarinia在那一年的冬天就会死亡，也就是说，观测到确认存在Vespa mandarinia的那个地方在政府完成除虫工作后虽然有可能仍然存在几只侥幸逃脱的Vespa mandarinia，但是这些Vespa mandarinia只能在那一年内存活，一旦度过冬天进入第二年的春天，这些Vespa mandarinia就会死亡。

当一份新的报告被提交，报告中会记录群众发现该物种时的地点与时间，即所提供的数据集中的“Latitude”和“Longitude”，如果这个地点与当年内被确认存在Vespa mandarinia的地点之间的距离非常接近，那么显然，这个物种是Vespa mandarinia的概率就越高，这份报告是错认的概率就越低。有次我们可以提取处一个新的特征变量，这个特征变量表示的就是每一份报告中发现一个可能是Vespa mandarinia的物种的地点与当年内被确认存在Vespa mandarinia的地点的距离，我们称其为“distinction”。特征变量“distinction”的计算过程如下。

我们将第一个被确认存在Vespa mandarinia的地点的“distinction”设为0，根据所提供的数据，这个地点是在2019/9/19被发现的，其“Latitude”为49.149394，其“Longitude”为-123.943134。根据这个点我们有如下的计算：

首先对于在2019/9/19之前呗提交的报告，因为在这个时间节点之前没有被确认的Vespa mandarinia，所以我们认为这些报告有很大的可能性是错分的，因此我们给这些报告的特征变量赋一个值M，M是一个无限趋近于正无穷的数。

其次，对于在这个时间节点之后提交的报告，我们首先根据报告中观测到那个可能是Vespa mandarinia的物种的时间，即“Detection Date”，提取出在这个时间节点之前并且在同一年内的所有被确认包含Vespa mandarinia的地点，然后根据经纬度与公里的转换公式：计算提交的报告中的地点到各个地点的距离，然后取这些距离距离的平均值作为这份报告的“distinction”。

经纬度与公里之间的转换公式如下所示：

D = arc cos((sin北纬A×sin北纬B) + bai(cos北纬A×cos北纬B×cosAB两地经度差))×地球平均半径(Shormin)

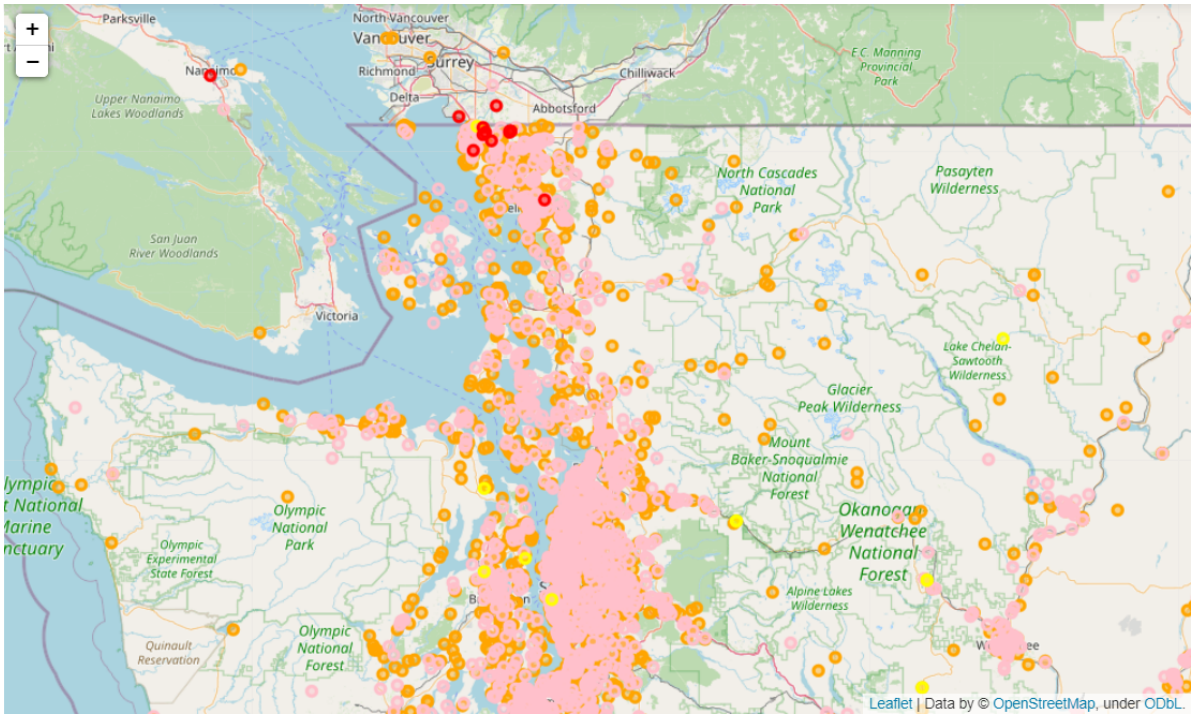
其中地球平均半径为6371.004 km，D的单位为km

四 蔓延情况预测

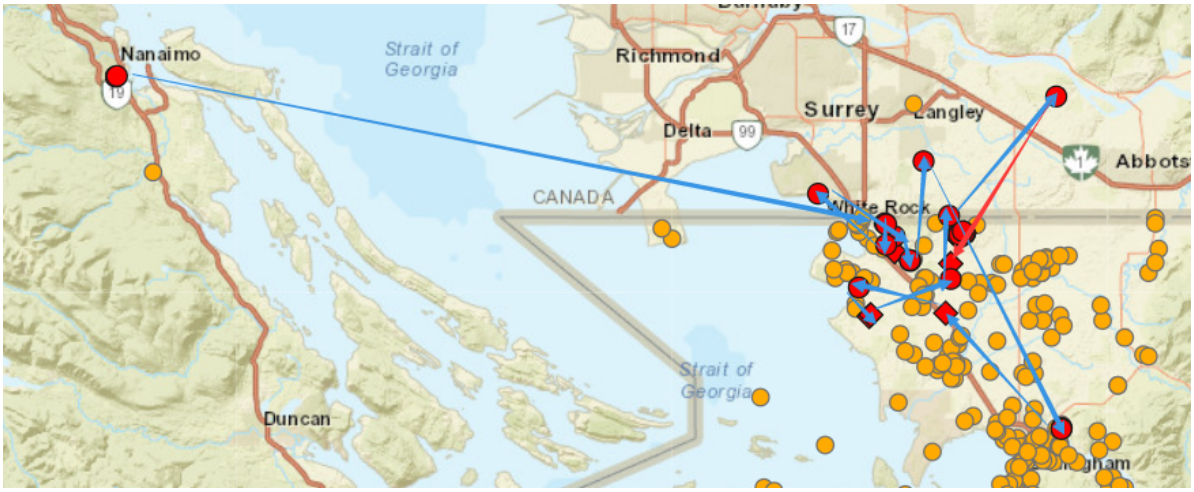
问题一要求我们预测Vespa mandarinia在一段时间内的蔓延情况并给出预测的精度。

根据所提供数据集中的"Longitude"和“Latitude”我们将所有的数据都绘制到地图上，绘制出的图像如下所示：

其中红色的点是被标志为“Positive ID”的报告的地点；粉色的点是被标志为“Negative ID”的报告的地点；橙色的点被标志为“Unverified”的报告的地点。我们发现被确认存在Vespa mandarinia的地区较为集中，大致分布在经度48到49，纬度-124到-122的区域内。其他提交的报告，包括错认的报告和无法判断的报告，则分布得较为分散，但是绝大部分也是位于水域边缘，这是因为根据虫类的习性，相较于内陆地区水域周围更容易生长虫类。



为了分析Vespa mandarinia的蔓延趋势，我们根据14个被标记为“Positive ID”的报告的发现的时间绘制出如下的蔓延图：

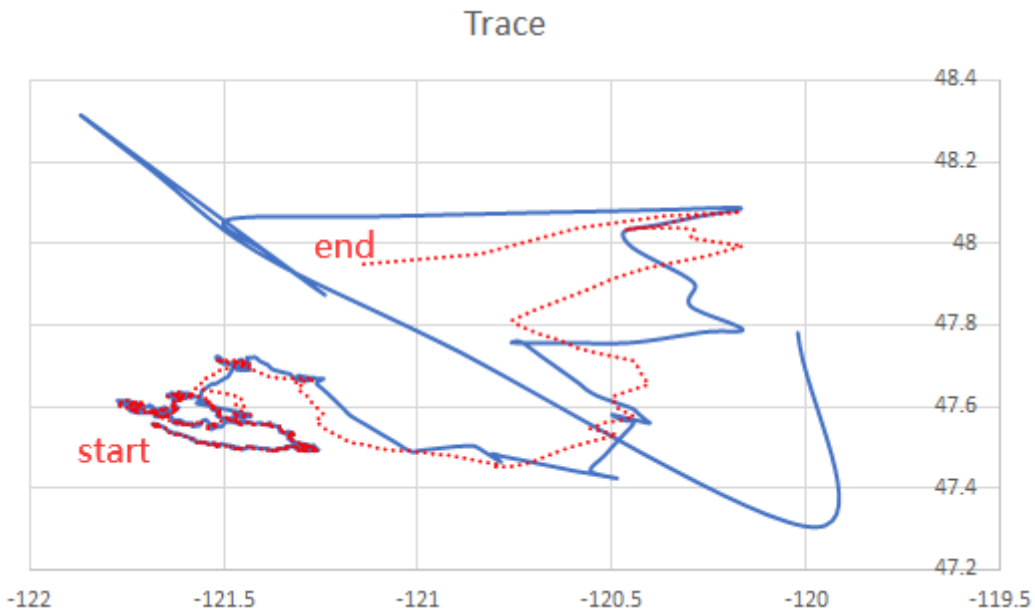


该图中的蓝色箭头表示14个Positive report的蔓延趋势，从这个图中我们可以发现Vespa mandarinia的分布有一种从里向外扩散，从水域向内陆扩散的趋势。

为了验证我们预测的趋势，我们希望能够利用时间序列模型准确的估计出Vespa mandarinia的蔓延。但是由于相关部门为我们提供的数据中仅有14条Positive 数据，并且从上图我们也可以看出这14条数据中还存在一个影响极大的离群点（左上角红点），我们并不能够利用Positive report的位置信息直接得出Vespa mandarinia的蔓延趋势。

我们运用逆向思维考虑这个问题，Vespa mandarinia会捕食其他蜂种，根据动物的天性，Vespa mandarinia会有一个向食物密集处迁移的趋势，即其他蜂种的迁移会引起Vespa mandarinia种群迁移的联动反应，并且Vespa mandarinia的迁移具有一定的滞后性。

由于其他蜂种是Vespa mandarinia的重要食物之一，而且属于同一个科目的物种更容易被混淆，所以那些Negative report中的物种很有可能是Vespa mandarinia的食物，Vespa mandarinia会随着他们的迁移二迁移。因此，通过预测Negative report中的地点信息的迁移我们可以得到Vespa mandarinia的大致迁移路径，如下图所示：



其中蓝色的线是Negative report中的物种迁移轨迹，红色的先是蓝线的滑动平均，我们可以将红线作为Vespa mandarinia的大致迁移路径，其迁移是从start点到end点。从这个图中我们可以发现Vespa mandarinia逐渐从密集变得分散，这与我们的预测是一致的，说明我们的预测较为准确，并且根据图中经纬度坐标的变换，我们还可以发现Vespa mandarinia有一个向东南方向蔓延的趋势。

综上，我们认为Vespa mandarinia有一种从里向外扩散的蔓延趋势，并且主要向东南方向扩散，经过我们的检验，这个预测的准确率较高。

五 逻辑回归模型

问题二要求我们利用所给的群众报告数据和图像文件建立一个用于预测其他昆虫被错认为是Vespa mandarinia的可能性的数学模型。

本题中的错认是指群众错误的将其他昆虫认为是Vespa mandarinia并向政府部分发送与该昆虫有关的报告的行为。错认在所给数据上的表示就是那些被专业机构判定为“Negative ID”的报告。本题中的可能性就是指其他昆虫被错认为是Vespa mandarinia的概率。我们认为由于没有专业人士的实地考察，每一份报告都有一个错认的概率，当这个概率高于一个阈值时，政府机构就可以判断这份报告是错认的并

给他一个“Negative ID”的标记；当低于这个阈值时我们就认为发送这份报告没有错认，并给这份报告一个“Positive ID”的标记。

综上，我们认为本题就是要构建一个数学模型，该模型能够得出每一篇报告的错认概率。之后通过模型拟合，我们能够得出一个概率阈值，并将高于这个阀值的报告定性为“Negative ID”，将低于这个概率阈值的报告定性为“Positive ID”。

考虑到最终我们需要的是一个概率值所以我们选择逻辑回归作为我们的模型

逻辑回归原理

逻辑回归模型是一种特殊的回归模型，其本质还是一个线性回归，只是他使用一个Logistic 函数将y值归一化，使得y值能够落在区间（0,1）内。逻辑回归的过程就是先将所有的特征变量线性相加，然后将结果带入Logitic函数g(z)中，从而可以得到逻辑回归的表达式，如下公式所示：

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

得到函数表达时候就可以利用参数估计的方法得到每个特征变量前的系数，从而得到一个真正的逻辑回归模型。

因为逻辑回归模型值的特殊性——分布在区间（0,1）内，我们可以将逻辑回归的结果看成是一种概率。

下一节是我们构建逻辑回归模型的过程，在构建的过程中我们使用最大似然估计得到模型的参数估计值。

构建逻辑回归模型

接下来我们构建逻辑回归模型并对其进行统计学检验。首先，因为政府部门为我们提供的数据存在严重的不平衡问题，所以我们采用重采样的方法使得被标记为“Positive ID”的数据与被标记为“Negative ID”的数据在数量上较为接近，最终我们得到一个平衡的数据集，这个数据集包含2069条被标记为“Negative ID”的数据和2058条被标记为“Positive ID”的数据。

之后，我们将数据预处理后得到的四个新特征“Notes相似度”，“图像相似度”，“time_probability”，“distinction”作为逻辑回归模型的输入，将每一分报告的“Lab_status”作为模型的输出，由此构建了一个二元逻辑回归模型。然后将70%的数据划分为训练集，30%的数据划分为测试集，再在训练集上进行模型的训练与拟合，在测试集上进行模型的检验。

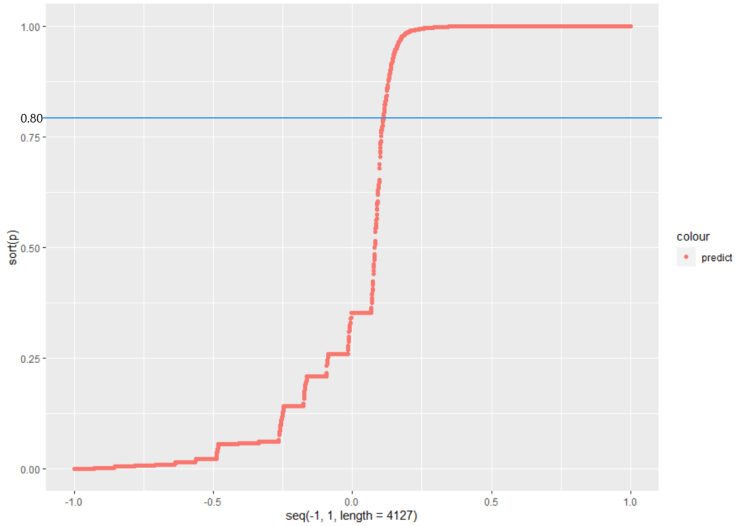
最终我们得到了如下公式所示的逻辑回归模型

$$\hat{tag} = \frac{1}{1+e^{-\beta}}$$

$$\beta = -18.152 + 1.948time + 20.981text + 1.672image + 71.481place$$

(0.806) (0.178) (1.929) (0.205) (3.501)

利用上述公式，我们可以计算每一篇报告被错认的概率，因为逻辑回归的输出值是一个概率，介于0和1之间。我们可以设定一个阈值，作为二分类划分的界限。对于Vespa mandarinia这一具体问题，经过我们的反复调参，我们最终确定阈值设置为0.8，此时逻辑回归对这个二分类问题的拟合程度最好，准确率为0.9583。每一分报告的错分概率如下图所示：



通过观察图 可以发现，数据较为均衡的分布在0.8这条水平线的上下两部分，这与我们模型的结果相一致。

至此，我们可以利用该模型高效的判断一份报告是否错认。如果这份报告的概率高于阈值0.8，那么这份报告就是错认的，会被标记为“Negative ID”；如果这份报告的的概率小于阈值0.8，那么这份报告就是正确的，会被标记为“Positive ID”。一旦出现一份被标记为“Positive ID”的报告，政府有关部门就必须重视起来。

六 逻辑回归模型统计学检验

因为逻辑回归是一个统计模型，我们需要对其进行一系列的统计检验，检验结果将在本节中展示，每节内容如下所示：

6.1节：假设检验

6.2节：区间估计

6.3节：拟合优度

6.4节：灵敏度分析

6.5节：ROC曲线

假设检验

假设检验就是构建一个原假设H0，假设原假设为真，然后看数据是否提供了反对H0的证据。在逻辑回归问题中，学术界一般构建z统计量来检验H0.

在本节中我们想要检验每一个特征变量是否对解释变量“Lab status”有影响，如果有影响，那么这些逻辑回归得出的这些特征变量的系数就不为0。所以我们做出如下假设：

H0：β_i=0

只要我们能够证明原假设不成立，那么就是有影响的，可以保留在我们的模型中。

在原假设下，z统计量的值就是回归系数除以其标准误差，公式如下所示：

$$Z = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} = \hat{\beta} / \hat{\sigma}_{\hat{\beta}}$$

在设定显著性水平为$\alpha = 0.05$时，根据经验，如果Z值的绝对值大于2.0，则该特征变量对解释变量有很重要的影响。

利用公式，我们可以得到每一个特征变量的z值，结果如下表所示：

z
10.96
10.88
8.16
20.42
-22.52

每一个z值得绝对值都远超2.0，所以这4个变量对Lab Status的预测结果具有很重要的影响，每一个变量都需要被保留在模型中。

区间估计

区间估计是在点估计的基础上，给出系数总体参数估计的一个区间范围，该区间通常由样本统计量加减估计误差得到。在统计学上这个区间又叫做置信区间。

在我们的模型中，每个变量系数的95%置信区间如下表所示：

[95% Conf.	Interval]
1.599367	2.296145
17.2001	24.7609
1.27031	2.073816
64.61973	78.34186
-19.73215	-16.57282

拟合优度

在一般的线性回归模型中有一个指标R^2用于表示模型的拟合效果，在逻辑回归中这个指标被称为$Pseudo-R^2$,这个指标有多种定义方式，最常见的一种就是McFadden's R2（似然比检验），其公式如下所示：

$$\mathrm{McFadden's\,R^2} = 1 - \frac{\ln \mathrm{L}}{\ln \mathrm{L_0}}$$

其中，L_0指仅包含常数项的模型的似然值，L指包含所有解释变量以及常数项的的似然值。

通常，$McFadden's R^2$的值只要大于0.5就认为模型具有较好的拟合效果。

除了$Pseudo-R^2$,逻辑回归模型也可以使用信息准则来评估模型的拟合度，常用的信息准则有AIC和BIC

AIC准则定义为

$$AIC = \frac{-2\ln \hat{L}(M_k) + 2P}{N}$$

其中$\hat{L}(M_k)$是模型的似然值，P是模型中参数的个数（包含常数项），在其他条件相同时，AIC的值越小，则模型拟合程度越高。

BIC准则定义如下

$$BIC = -2\ln L(M_{Full}) - (N - P)\ln N$$

同样，BIC值越小表明相应模型拟合度越高。

在外面的检验中这三种定义的计算结果如下表所示：

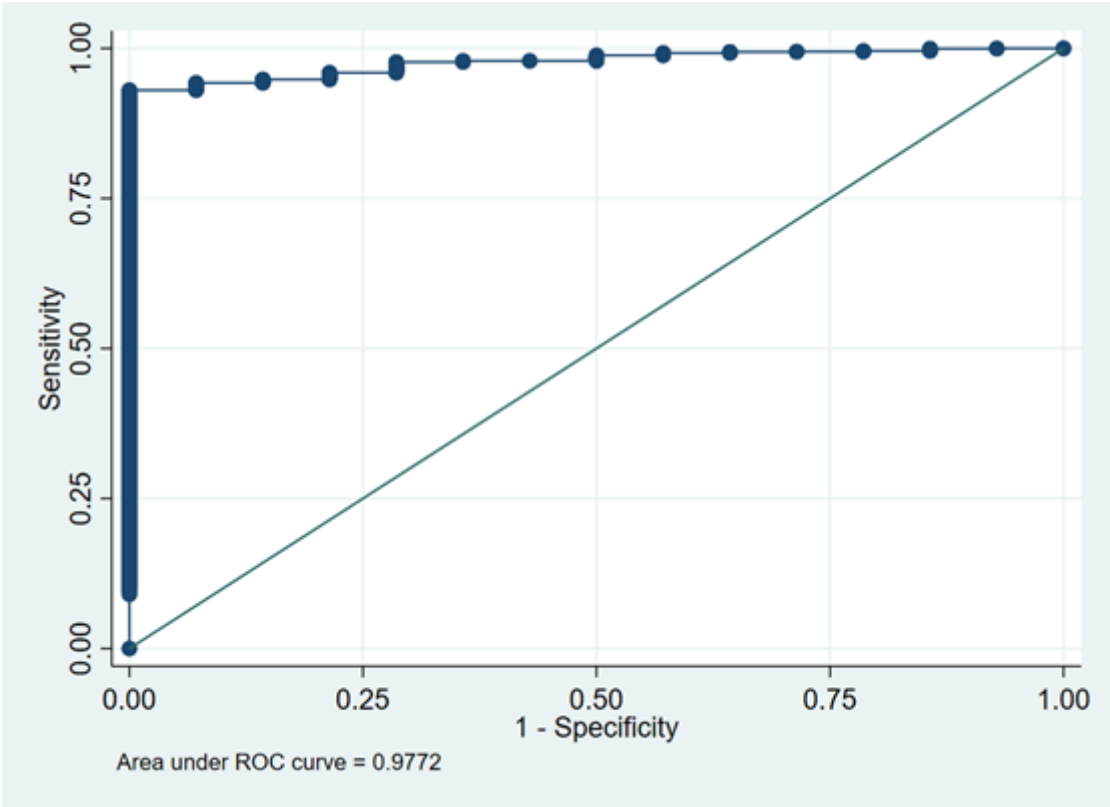
```
. fitstat

Measures of Fit for logit of tag

Log-Lik Intercept Only:     -2860.604     Log-Lik Full Model:          -626.158
D(4122):                     1252.317     LR(4):                      4468.891
                                          Prob > LR:                     0.000
McFadden's R2:                  0.781     McFadden's Adj R2:             0.779
Maximum Likelihood R2:          1.000     Cragg & Uhler's R2:            1.000
McKelvey and Zavoina's R2:      0.981     Efron's R2:                    0.839
Variance of y*:               173.874     Variance of error:             3.290
Count R2:                       0.958     Adj Count R2:                  0.916
AIC:                            0.306     AIC*n:                      1262.317
BIC:                       -33064.595     BIC':                      -4435.590
```

我们可以看出模型的拟合结果很好。

灵敏度分析

ROC曲线分析

ROC曲线将灵敏度与特异性以图示方法结合在一起，ROC曲线越凸即越近左上角则表明这个模型的价值越大，模型效果越好。我们的ROC曲线如下图所示：



可以看出ROC曲线非常凸，说明我们的模型效果很好。

七 正面目击报告筛选

问题三要求我们使用所建立的模型，筛选最有可能的正面目击报告进行调查。

我们在问题二种建立的逻辑回归模型能够根据输入的四个特征“Notes_similarity”，“Image_similarity”，“time_probability”和“distinction”得出一份报告是错认的概率，当这个概率高于0.8时，这份报告就是错认的，会被标记上“Negative ID”，否则这份报告就是正确的，会被标记上“Positive ID‘。显然，逻辑回归模型输出的概率越大这份报告就越可能是错认的；输出的概率越小，这份报告的正确性就越高，也就是说这份报告所记载的地点越有可能存在Vespa mandarinia，当地政府就需要进行除虫工作。

所以，为了解决问题三，我们可以计算出每一份提交的报告的错认概率，然后进行按照从小到到的顺序排列这些报告，位置越靠前的数据就越有可能是正面目击报告。

为了验证我们的筛选方法的有效性，我们计算出已经有专业人士标记的2083分报告的错分概率，排序后发现被标记为”Positive ID“的14份报告均位于顶部，部分被标记为”Positive ID“的报告的概率如下表所示：

GlobalID p

{5AC8034E-5B46-4294-85F0-5B13117EBEFE} 0.0021

{5EAD3364-2CA7-4A39-9A53-7F9DCF5D2041} 0.5151

{124B9BFA-7F7B-4B8E-8A56-42E067F0F72E} 0.6188

{AD56E8D0-CC43-45B5-B042-94D1712322B9} 0.6828

{F1864CC3-508C-4E60-9098-B158AB413B03} 0.7124

这个结果有力的验证了我们的筛选方法，之后我们将我们的筛选方法用于被标记为”Unverfied“和”Unprocessed“报告，得到了前十份最有可能是正面目击报告的数据，如下表所示

第几 GlobalID p

八 模型更新

问题四要求我们提供一种随着时间的推移更新我们的模型的方法，并给出跟心模型的频率。

在构建逻辑回归模型时我们给这个模型四个输入特征：“”Notes_similarity“”，“Image_similarity”，“time_probability”,“distinction”。并使该模型能够输出每一份报告是错分的概率。由于不断的有群众提交自己的report，我们的模型就需要根据群众提交的报告更新。因为我们在建模的最终目的是希望快速高效的识别出是Positive的报告，再加上群众提交的report绝大部分是Negative的，仅有极少的积分报告试试Positive的，所以我们的模型更新应该由新检测到是“Positive ID”的报告所决定。经过我们的分析，我们得出每当检测到一个新的被标记为“positive ID”的报告时，模型就会更新。模型更新的方式如下。

群众新提交的report中是否附带图像文件或者是否给出“Notes”都是可选的，但根据描述性统计的结果，大多数报告都会包含这两个内容。所以随着检测到的Negative report的数量的增加，基于Ngative report中“Lab comments”特征变量的易混淆特征关键词表会被扩充，Negative图象也会增加。因此计算“Notes_similarity”和“Image_similarity”的子模型——TF-IDF模型和卷积神经网络模型就会更新。

另外群众提交的报告中一般包含“Detection Date”， “Latitude” 和 “Longitude” 这三个特征变量。这三个变量可以产生输入逻辑回归模型的另外两个变量“time_probability”和“distinction”。但这两个模型的计算方式并不会因为新增加的Negative report而改变。“time_probability”需要计算每个月对应的可能会出现Vespa mandarinia的概率，“distinction”需要寻找同一年内的Positive report的位置，所以当增

加一份被标记为“Positive ID”的报告时，这两个特征变量的计算方式会发生改变。

因为从灵敏度分析结果来看，“time_probability”和“distinction”相较于另外两个特征并不稳定，数据的小变动就是引起逻辑回归模型中这两个特征变量的大变动，但并不会引起“Notes_similarity”和“Image_similarity”的大变化。同时考虑到群众会频繁提交Negative report，每次都更新会严重浪费资源，所以我们认为只有检测到PositiveID 时，为了后续预测的精确度必须需要调整模型参数，更新模型。“time_probability”和“distinction”在计算时需要将新的Positive report的的相关信息加入计算中。同时更新易混淆特征关键词表和Negative 图像的数据更新的内容是两次模型更新的时间间隔内累计的negative report 的“Lab comments” 数据和Negative 图象数据。进而更新逻辑回归的各项回归系数。

九 问题五模型构建

问题五要求我们利用自己的模型给出一个可以证明Vespa mandarinia已经完全被消灭的证据。

根据文献资料，我们知道只有受精的皇后才能生育其他的Vespa mandarinia，并且到了冬天除了皇后，其他的Vespa mandarinia都会死亡。因此，我们认为证明Vespa mandarinia被根除的理想证据是下一年不会再有任何一只蜂后生育或者说本年度已经将所有蜂后消灭。

我们的模型可以预测报告是错认的概率，从而高效的提取出每一份Positive report，每一份Positive report正好对应一个模型更新的时间节点。假设最后一次更新模型的时间节点为T_N，对应的Positive report中的Vespa mandarinia就是到当前时间节点为止的最后一只被观测到的Vespa mandarinia，这个Vespa mandarinia可能是蜂后也可能是普通的Vespa mandarinia。假设同一年度内T_N节点之前的每个更新节点为T1,T2,T3,....T_N-1。T_i之间的间隔为△T1,△T2,△T3,....,△T_N-1。

当$\lim_{T_N->T_i}(1/\triangle T)=0$时，即当当前时间节点$T_N$趋向于某一个特定的时间节点$T_m$时，$1/\triangle T$趋向于0就可以认为是Vespa mandarinia已经都被消灭。另一种解释就是在某一个时间节点T_i附近，发现一只新的Vespa mandarinia需要的时间间隔为正无穷，那么发现一只新的蜂后的时间间隔一定也是正无穷，那么就说明本年度内的所有蜂后已经全部被消灭。

模型优缺点

模型优点：

我们的模型综合利用时间、空间、文本和图象数据，多投票机制增加了模型的鲁棒性。

我们的模型可扩展性强，很容易在新数据集上实现，并且易于更新

模型缺点：

我们的模型可能忽略了其他一些对Vespa mandarinia分布具有重要影响的变量，比如温度。

十 备忘录

针对相关单位提出的五个问题，我们做如下总结：

- Address and discuss whether or not the spread of this pest over time can be predicted, and with what level of precision
  我们提出的模型是基于提交的report信息并结合十分微小的实际传播路径信息进行预测的。利用已知的数据将所有打了标签的report画在地图上可以看到2019年至今，大黄蜂的分布。受限于数据的丰富性，我们利用已知的为数不多的positiveID样本模拟出了实际的蜂群数量，这样就弥补了数据量匮乏的不足，再结合位置时间数据得出了Vespa mandarinia种群蔓延的大致趋势。模型指出Vespa mandarinia逐渐向华盛顿州东南部蔓延，需要提醒居民们提前注意防范。
- Most reported sightings mistake other hornets for the Vespa mandarinia. Use only the data set file provided, and (possibly) the image files provided, to create, analyze, and discuss a model that predicts the likelihood of a mistaken classification.
  大多数report都是被判断为negativeID的，为了解决这个烦恼，我们模型的目标是，从提交的

众多份report中优先选出最有可能为PositiveID的来给专家校验，节省了大部分被negativeID report耽误的时间。所以为了提高预测的准确率，我们对提供的数据进行了具体的分析，包括文本数据，图像数据，时间数据，地理位置数据。由于数据具有严重的不平衡问题，所以并没有直接利用这四个特征，而是在最终的模型之前加入子模型提取特征。文本数据特征的处理使用了tf-idf模型，图像数据使用了resNext50模型，时间数据和地理位置数据处理成频率数据，分别是时空上与negativeID样本的相似性。把这4个子模型的输出看作一份report属于negativeID与positiveID的投票数，将该投票作为逻辑回归的输入，即可得到预测的标签。

- Use your model to discuss how your classification analyses leads to prioritizing investigation of the reports most likely to be positive sightings
  逻辑回归模型会输出属于negativeID的概率，经过一番调节参数，模型可以将数据集中属于positiveID的样本都预测出来，在此基础上对unprocessed和unverfied的样本做出同样的预测，模型给出如下两个很可能也是Vespa mandarinia。
  {9A5CB940-8951-4FE7-8619-DAF4A8FE1850}
  {41EB9554-6676-4CD1-BE5B-5133CB3EE4B0}
- Address how you could update your model given additional new reports over time, and how often the updates should occur
  特征包括：
  时间：report所在月份，领取历史数据中negativeID出现的月份的比率。
  空间：report所在地理位置与目前所有positiveID的距离的最大值，然后作概率处理。
  文本：report notes与目前所有negativeID 的词典的相似度。
  图象：report image与目前所有negativeID图象的相似度。
  由于提交的report 大部分都是negativeID ,所以更新模型的频率由实际被检测到positiveID决定，即，当检测到一个新的positiveID 时，就会更新模型。更新模型的方法如下：
  新的report文本和图象是可选的，但大多数都会包含这两个内容。时间和空间是必然包括的，也是在基于历史数据预测未来情况的情景下对模型来说最优的数据字段。随着检测到的negativeID的数量的增加，基于negativeID的文本相似度计算的字典会被扩充，图象的负面素材也会增加。但我们把模型更新的时间点定在下一次检测到PositiveID时，因为检测到PositiveID时，会引起时间特征和空间特征的变化，从灵敏度分析来看这两个特征的参数，并不是十分的稳定，那么在检测到PositiveID 时，为了后续预测的精确度需要调整模型参数。即，时间频率特征和空间特征需要将新的PositiveID的相关时间和地理位置加入频率的计算中，同时扩充一下另外两个特征的数据：两次模型更新的时间间隔内累计的negativeID 的文本素材和图象素材。进而更新逻辑回归的各项回归系数。
- Using your model, what would constitute evidence that the pest has been eradicated in Washington State?
  判断这种害虫被根除的理想状态是下一年不会再有任何一只蜂后产卵，或者说本年度已经将所有蜂后根除，因为其他胡黄蜂不会活过一年以上，而且不具有繁殖的能力。时间加空间的特征组合可以定位蜂后的出没，恰好对应了模型更新的时间点。
  根据实际的模型更新时间节点序列的分布，建立检验统计量即可得出判断。

最后，为了使模型更好的工作，模型假设要保持尽量的满足的状态，否则模型可能失效。