**《数据科学与数据分析》课程作业**

**专业：** 信息管理学院 **学号：** 211820073 **姓名：** 胡涂

**专业：** 信息管理学院 **学号：** 211840182 **姓名：** 邓雨茵

**专业：** 信息管理学院 **学号：** 211820274 **姓名：** 郑宇

# 在线文章热度的冷启动识别

## 介绍

### 研究背景

在线文章，例如各类社交平台以及新闻平台上的博客以及文章，已经成为现今最重要的信息来源。在线文章的热度，即受欢迎程度，可以用文章的转发量来衡量。

现有的研究中，多数对在线新闻文章的预测，是采用文章时间过程特征以及文本内容特征融合，混合预测高热度文章和与低热度文章。如将时间过程与内容特征进行深度融合对微信公众号文章的浏览量进行预测(Liao et al., 2019)，在不平衡测试集中对热度分类的预测准确率达到了0.965。

这样的预测方法在工业应用中的价值很高，同时预测准确率也较高。然而对于内容创作者而言会产生一些问题，如高热度文章具有极大的偶然性，对于创作者来说很难从中得到启示，并将经验应用到自己文章之中；很难在无法获得时间过程特征的情况下保持较高的预测准确率。

这样的问题与解决文章的“冷启动问题”的描述相符。文章的“冷启动问题”(Arapakis, Cambazoglu, & Lalmas, 2014)，即在发表前预测文章热度的方法，现在仍旧在起步阶段。过往研究认为，文章的发布平台是文章热度的决定性因素。如果内容创作者想要调整文章的结构等特征，这将是没有帮助的。同时对冷启动问题，也有研究基于文章标题特征(Piotrkowicz, Dimitrova, Otterbacher, & Markert, 2017)，进行标题的特征建模，从而对新闻热度进行回归。在他们的研究中，对文章热度的回归模型MAE值达到了0.67，这是在对响应变量进行对数转换得出的。本文认为，这样的误差是不能接受的。

同时，过往所有的研究都将低热度和高热度合并当作研究重点。Arapakis等人的研究中指出，对文章热度分箱处理转分类任务，这样的操作会对冷门文章产生偏见。然而，冷启动回归任务的效果并不能达到预期的效果，噪声巨大。

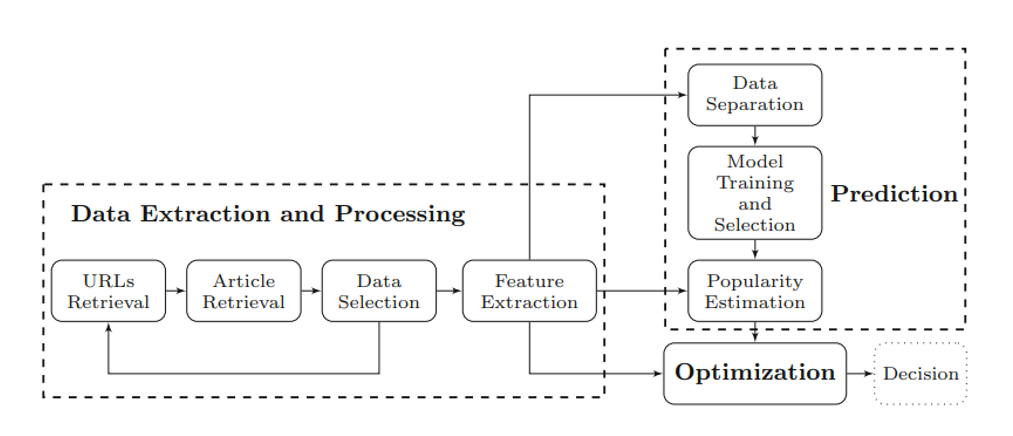
### 任务介绍

前面提到，高热度文章的偶然性与热点的强关联性很难让内容创作者得到有用的启示。

而极低热度文章，它们或许具有一些令创作者感兴趣的内容特征，能够使文章避免低热度，使得文章能够更关注于内容创作。

为了能够更好地向内容创作者提出建议，便于文章摆脱低热度，能够让内容获取者更注重文章内容。我们采取手工制作文章特征的方式，基于文章文本特征识别低热度文章与正常文章的任务，同时对文本特征进行相应解释，以便向内容创作者提出修改建议。

为了解决这一问题，我们将采用数据获取-特征构建-模型拟合-模型解释的流程。流程如下：



## 数据获取与特征构建

数据字典是对数据集中的各个特征进行解释的文档，它可以帮助我们理解数据的含义和来源。本文根据不同的特征类型，将数据字典分为以下几个部分：

（一）数据集获取

本研究使用的数据集来源于https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity，该数据集包含了两年内由Mashable发布的文章的各种特征和社交媒体分享数。原数据集共有39797条记录和61个特征，本文根据删除了没有明确语义含义或有歧义的特征，并根据过往研究与讨论，从文本中提取若干特征。

（二）特征的添加

我们利用Python网络爬虫技术根据数据集中提供的网址，爬取了文章的标题和正文内容，并且将其转化为文本格式。同时，考虑到新闻文章的热度与其所报道的事物的热度密切相关，我们还爬取了文章所属主题在文章发布前一周内的谷歌趋势指数。

### 1、时间特征

文章发布时间会影响文章的新颖性和时效性，从而影响读者的兴趣和分享意愿。一般来说，对于同一个热点事物，越早发布的文章越能抓住读者的注意力，越容易被分享。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| timedelta | 距离收集时间的间隔，以天为单位。 | Number |
| Date | 集数据的日期，格式为年-月-日。 | Date |
| dayRatio | 文章发表的前一天相较于前两天的热度增长率 | Number(0~1) |
| threeDayRatio | 文章发表的前三天相较于前六到三天的热度的增长率 | Number(0~1) |
| weekRatio | 文章发表的前一周相较于前两周的热度的增长率 | Number(0~1) |
| twoWeekRatio | 文章发表的前两周相较于前四到二的热度的增长率 | Number(0~1) |

文章的发布时间可能与文章的主题和内容相关，从而影响读者的偏好和分享动机例如，一些文章可能与特定的节假日或事件有关，如果在适当的时间发布，可能会引起更多的共鸣和分享。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| weekday\_is\_monday | 是否是星期一，是为1，否为0(下同)。 | Factor(0,1) |
| weekday\_is\_tuesday | 是否是星期二 | Factor(0,1) |
| weekday\_is\_wednesday | 是否是星期三 | Factor(0,1) |
| weekday\_is\_thursday | 是否是星期四 | Factor(0,1) |
| weekday\_is\_friday | 是否是星期五 | Factor(0,1) |
| weekday\_is\_saturday | 是否是星期六 | Factor(0,1) |
| weekday\_is\_sunday | 是否是星期日 | Factor(0,1) |
| is\_weekend | 是否是周末 | Factor(0,1) |
| isHoliday | 是否是节假日 | Factor(0,1) |
| HolidayName | 如果是节假日，节假日名称（Not Holiday表示不是节假日）。 | Factor |
| HolidayDay | 如果是节假日，是节假日的第几天（从1开始计数）。 | Number |
| HolidayDaysLeft | 如果是节假日，距离节假日结束还有几天（最后一天算1）。 | Number |

### 2、栏目特征

栏目可能影响文章的主题和内容，从而影响读者的兴趣和偏好，以及分享意愿。一般来说，不同的栏目可能吸引不同的读者群体，有些栏目可能更受欢迎或更有争议，从而更容易被分享。(李泽月, 2019)

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| data\_channel\_is\_lifestyle | 是否属于生活方式栏目，是为1，否为0 | Factor(0,1) |
| data\_channel\_is\_entertainment | 是否属于娱乐栏目 | Factor(0,1) |
| data\_channel\_is\_bus | 是否属于商业栏目 | Factor(0,1) |
| data\_channel\_is\_socmed | 是否属于社交媒体栏目 | Factor(0,1) |
| data\_channel\_is\_tech | 是否属于科技栏目 | Factor(0,1) |
| data\_channel\_is\_world | 是否属于世界栏目 | Factor(0,1) |

### 3、标题特征

标题字数、比较级与最高级词数、强化词与弱化词数、标点符号个数、不同词性的词数、标题主观性和标题情感极性会影响文章的吸引力和刺激性，从而影响读者的点击率和分享意愿。一般来说，文章的吸引力和刺激性越强，文章越能激发读者的好奇心和情感，越容易被分享，但也可能被认为是哗众取宠的表现而被忽视。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| n\_tokens\_title | 标题字数，以空格分隔的单词数 | Number |
| Comparatives Count | 标题比较级个数 | Number |
| Superlatives Count | 标题最高级个数 | Number |
| Count Intensifiers | 标题强化词个数 | Number |
| Count Downtoners | 标题弱化词个数 | Number |
| noun\_count | 标题名词个数 | Number |
| verb\_count | 标题动词个数 | Number |
| adverb\_count | 标题副词个数 | Number |
| punc\_count | 标题标点符号个数 | Number |
| title\_subjectivity | 标题主观性，越高表示标题越带有个人情感或观点 | Number(0~1) |
| title\_sentiment\_polarity | 标题情感极性，越高表示标题越积极或正面。 | Number(0~1) |

标题的可读性、简洁性、准确性和新奇性对于吸引读者和促进分享非常重要。本研究探讨了四个影响标题可理解性的特征：可读性指数、语法分析树高度、歧义和新奇程度。

可读性指数反映了标题的阅读难度，我们使用 Flesch-Kincaid Grade Level 公式来计算。

语法分析树高度反映了标题的结构复杂度，我们使用Python中的NLTK中的句法分析器来生成和分析。

歧义反映了标题的表达清晰度，我们使用词语含义数的乘积来估计。

新奇程度我们使用当前标题tf-idf的和在它之前的标题的tf-idf余弦相似度最大值得到。

我们发现，这些特征越低，标题越新奇，容易阅读和分享；但是如果可读性、简洁性、准确性过低，标题可能会缺乏复杂性和深度，从而降低其价值和信任度。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| **Flesch Kincaid Grade of Title** | 标题可读性指数 | Number |
| **SyntaxTree Height** | 标题语法分析树高度 | Number |
| All Possible Meanings | 标题所有可能意思（歧义） | Number |
| novel of title | 新奇程度 | Number(0~1) |

### 4、正文特征

文章的内容丰富程度是指文章能够提供多少有价值和有意义的信息给读者，从而满足读者的需求和期望。

文章的正文字数、不同字数占比、非停用词占比和非停用词不同字数占比反映了文章的长度和信息量。

外链和自链可以增加文章的信息来源和参考资料，扩展读者的知识面和视野，从而提高文章的内容丰富程度。

图片和视频可以增加文章的视觉效果和趣味性，丰富读者的感官体验和情感反应，从而提高文章的内容丰富程度。文章中名词，形容词和动词的占比反映了文章的语言风格。

名词占比较高的文章可能更客观和专业，形容词占比较高的文章可能更主观和描述性，动词占比较高的文章可能更动态和有说服力。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| n\_tokens\_content | 正文字数 | Number |
| n\_unique\_tokens | 正文不重复的单词数 | Number |
| n\_non\_stop\_words | 正文非停用词数 | Number |
| n\_non\_stop\_words | 正文非停用不重复的单词数 | Number |
| num\_hrefs | 外链个数，指向其他网站的超链接数。 | Number |
| num\_self\_hrefs | 自链个数，指向本网站的超链接数。 | Number |
| num\_imgs | 图片个数 | Number |
| num\_videos | 视频个数 | Number(0~1) |
| NWordRatio | 名词占比 | Number(0~1) |
| JWordRatio | 形容词占比 | Number(0~1) |
| VWordRatio | 动词占比 | Number(0~1) |

文章的主观程度和情感极性是指文章在多大程度上反映作者的个人看法和态度，以及作者对于文章主题的正面或负面的评价。主观程度越高，文章越倾向于表达作者的个人观点和情感，主观程度越低，文章越倾向于表达客观事实和数据。情感极性越高，文章越倾向于表达正面的情感，情感极性越低，文章越倾向于表达负面的情感。文章的主观程度和情感极性可以影响读者对文章的认同和共鸣，从而影响文章的传播效果。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| global\_subjectivity | 文章的主观程度(越大越主观) | Number(0~1) |
| globalsentimentpolarity | 文章的情感极性(正代表积极，负代表消极，绝对值越大情感越强烈) | Number(-1~1) |
| globalratepositive\_words | 积极词的占比 | Number(0~1) |
| globalratenegative\_words | 消极词的占比 | Number(0~1) |
| ratepositivewords | 非中性词积极词的占比 | Number(0~1) |

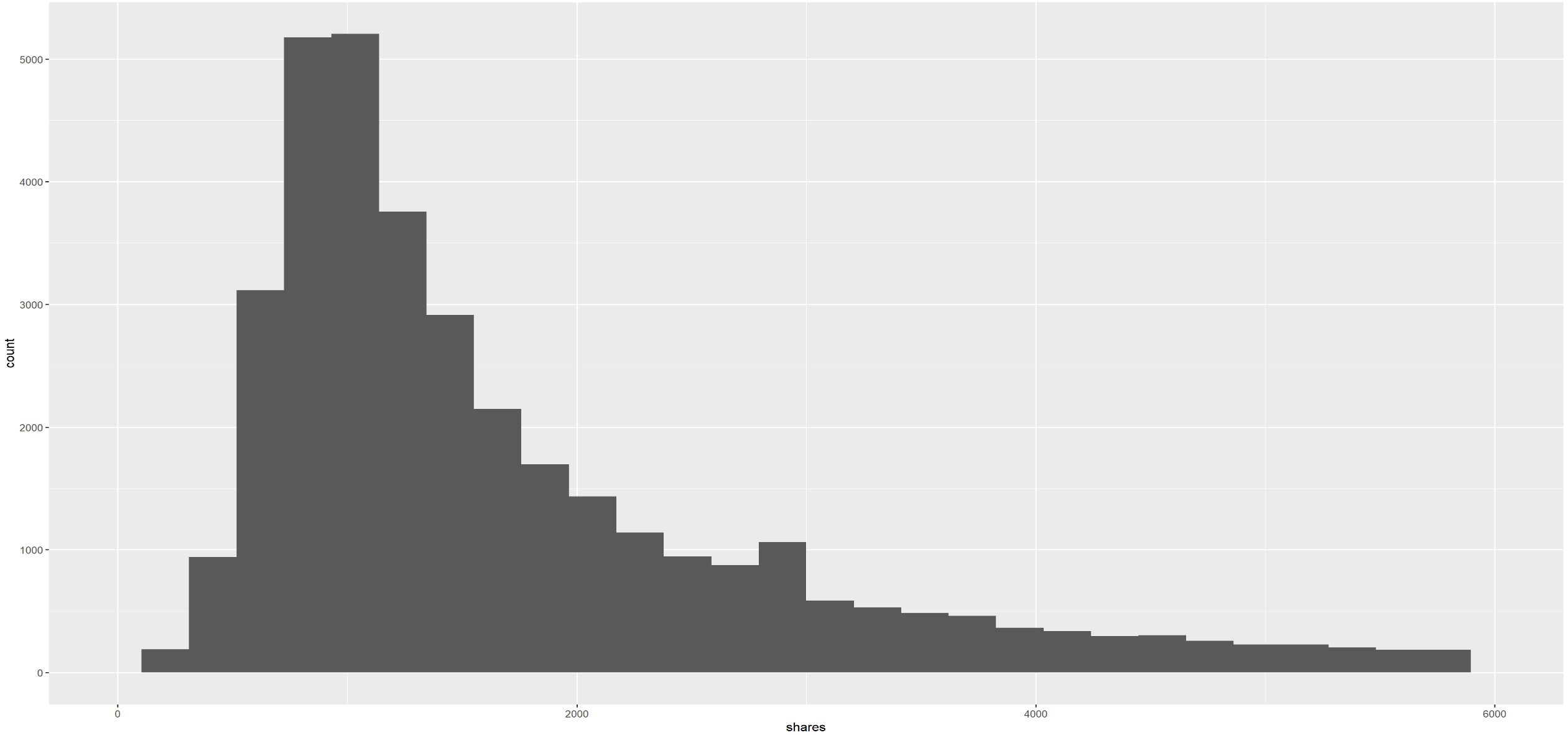
文章的可读性是指文章有多容易被读者理解，我们使用FleschReadingEase公式和文章所用词在牛津8000词的词的占比来衡量这一指标。文章越简单和清晰，可读性越低，文章越复杂和晦涩。

|  |  |  |
| --- | --- | --- |
| 特征 | 含义 | 属性 |
| ContentFleschReadingEase | 可读性 | Number |
| wordRatioIn8000 | 在牛津8000词的词的占比 | Number |

### 5、预测目标

shares： 文章被分享次数。

在此数据集中，shares分布是有一极长拖尾的偏态分布，分布直方图如下。



最小值为1，最大值为843300。5%分位点为584，95%分位点为10800。

在我们对shares分布的探索过程中，发现shares的分布与过往研究对于在线文章热度的分布的先验知识并不相同。在RNN+LSTM的研究中认为，冷门微信公众号文章（浏览量<100）的占比最多，达到了93%。而本文认为，Mashable这样的博客文章平台，冷门文章是占比较少的（5%）。

在Arapakis等人的研究中也提出，若采用冷门文章占多数这样的连续热度值分箱方法，会对冷门的在线文章形成一定的“偏见”。

因此，在后续的模型拟合过程中，我们采用了占比较少的冷门文章识别作为我们冷启动在线文章热度识别的主要任务。

## 模型拟合与优化过程

对于代表热度的shares连续值，本文采用了线性回归探索。并且对于我们指定的冷门文章识别任务，转为回归任务，利用采样+随机森林、SVM以及修改损失函数的神经网络进行模型拟合。部分采用了交叉验证，对采样类别和参数、模型超参数参数进行了调整，选择了最优参数。

### 回归任务与分析

对shares进行对数转换形成logShares，将logShares与截距项的线性回归模型作为基准模型，将前文所述所有特征对logShares拟合结果作为实验模型。回归结果如下。

相较于基准模型，所有特征对于响应变量在MAE的损失函数上有提升，证明特征是有效果的。然而，与Piotrkowicz等人的研究具有相同的问题。虽然将响应变量进行对数转换后会对回归任务的结果有着较好的提升效果，但是在真实值上仍然具有较大偏差。因为

将MAE的含义视作平均到每一个观测样本的偏差，可以不严谨的得到真实值的差距。

实验模型在真实值与预测值上的差距达到了2倍，本文认为这样的差距对于在线文章热度识别的任务是不可接受的。

本文猜测，导致损失函数差距过大的原因可能是：热度较大的在线文章具有很强的偶然性，噪声干扰极强，本文构建的内容特征具有较差的解释性。

因此，为了消除噪声的影响，同时较为契合的完成冷门文章识别的任务，本文将在后续研究中对shares分箱，进行分类任务。

### 分类任务

#### shares分箱处理

根据前文对shares分布的陈述，根据shares的5%分位点对其进行二分，记冷门为1，正常为2。同时根据shares的95%分位点对其进行二分，记正常为1，热门为2。（热门分类任务置于附录）。

由附录中的热门文章分类来看，本文构建的特征并不能很好地完成这一任务，噪声影响很大。因此，本文主要针对冷门文章识别任务做了尝试、改进与优化。

### 分类结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Recall-Cold | Recall-Normal | AUC |
| NaiveBayes | 0.792 | 0.341 | 0.815 | 0.64 |
| 采样+RandomForest | 0.705 | 0.64 | 0.709 | 0.746 |
| SVM | 0.609 | 0.682 | 0.605 | 0.696 |
| Focal Loss + 神经网络 | 0.722 | 0.751 | 0.188 | 0.522 |
| nnet | 0.575 | 0.75 | 0.566 | 0.724 |

### 模型优化

模型以朴素贝叶斯分类器（AUC=0.640）为基准线进行比较，根据分类器类别差异根据各自的性能评判标准进行优化。

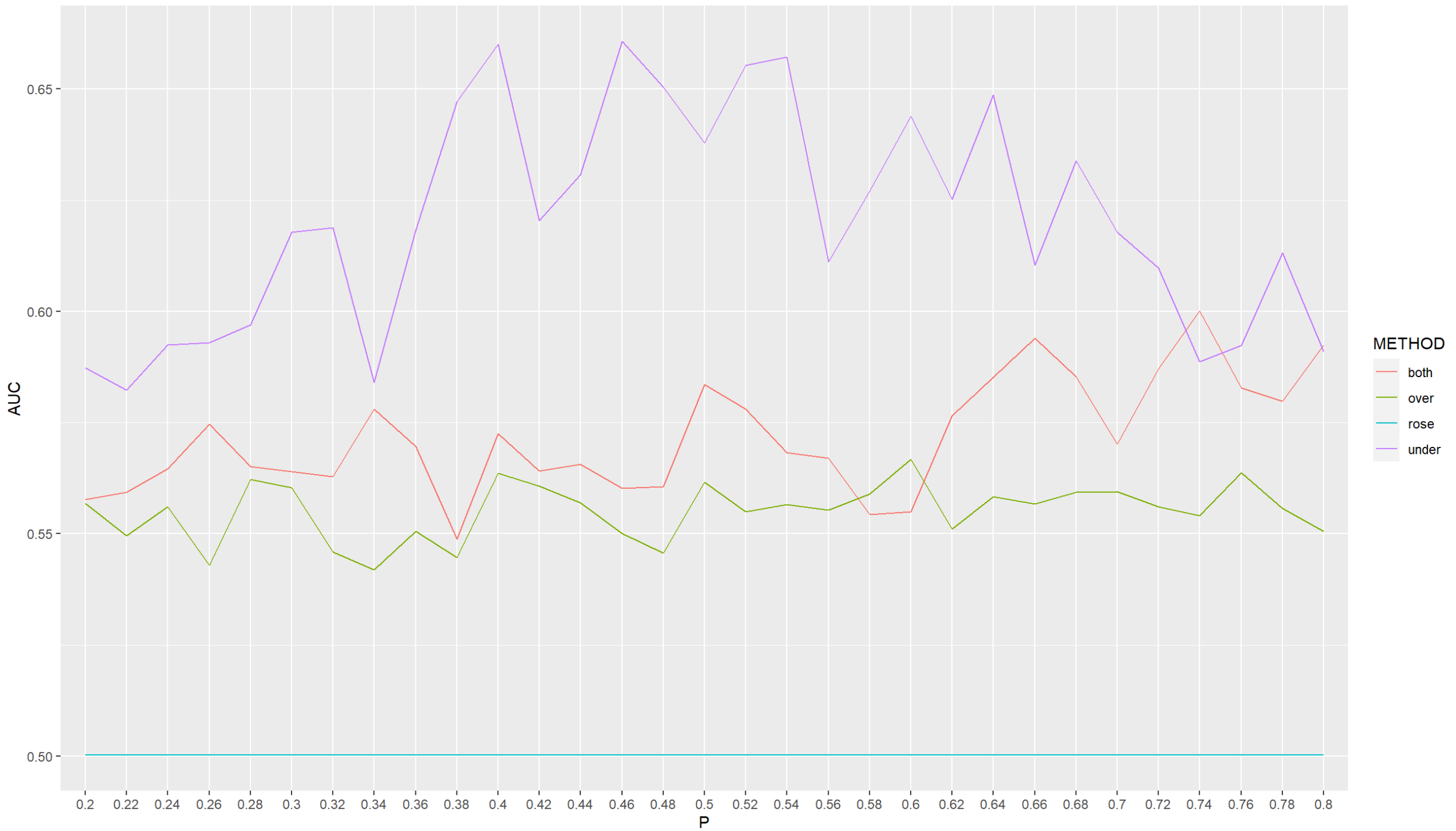
##### 2.3.1 随机森林优化

由于在类别标签中，样本不平衡程度较高(1:19)，可能会导致模型对于小样本特征学习不够，因此在随机森林模型拟合的过程中，采用测试集平衡采样+随机森林进行拟合。

在ROSE包提供了四种采样方法，对大样本的欠采样、对小样本的过采样、同时使用欠采样和过采样的双向采样以及基于原始数据的人工数据合成方法，同时方法中伴有参数p，代表小样本在采样后在总体中的占比。

在此处优化过程中，本文认为AUC值是最优的评估指标，因为它能排除不平衡样本所导致的影响。理想AUC值为1.0，最差AUC值为0.5，0.5代表着一个随机分类器。

因此，在模型拟合前，本文针对采样方法以及超参数p(0.2-0.8)进行了遍历搜索，并采用C5.0在采样后数据集上进行拟合，用测试集中的AUC值对最优参数进行选择，搜索结果如下。



可以看出，对于该数据集，欠采样+p=0.46附近的效果比较好，因此采样方法选择欠采样，p值选择0.46。最终训练数据集共有1755个观测样本。

作为对比，在使用平衡样本进行训练时，同时使用非平衡训练集对模型进行拟合。

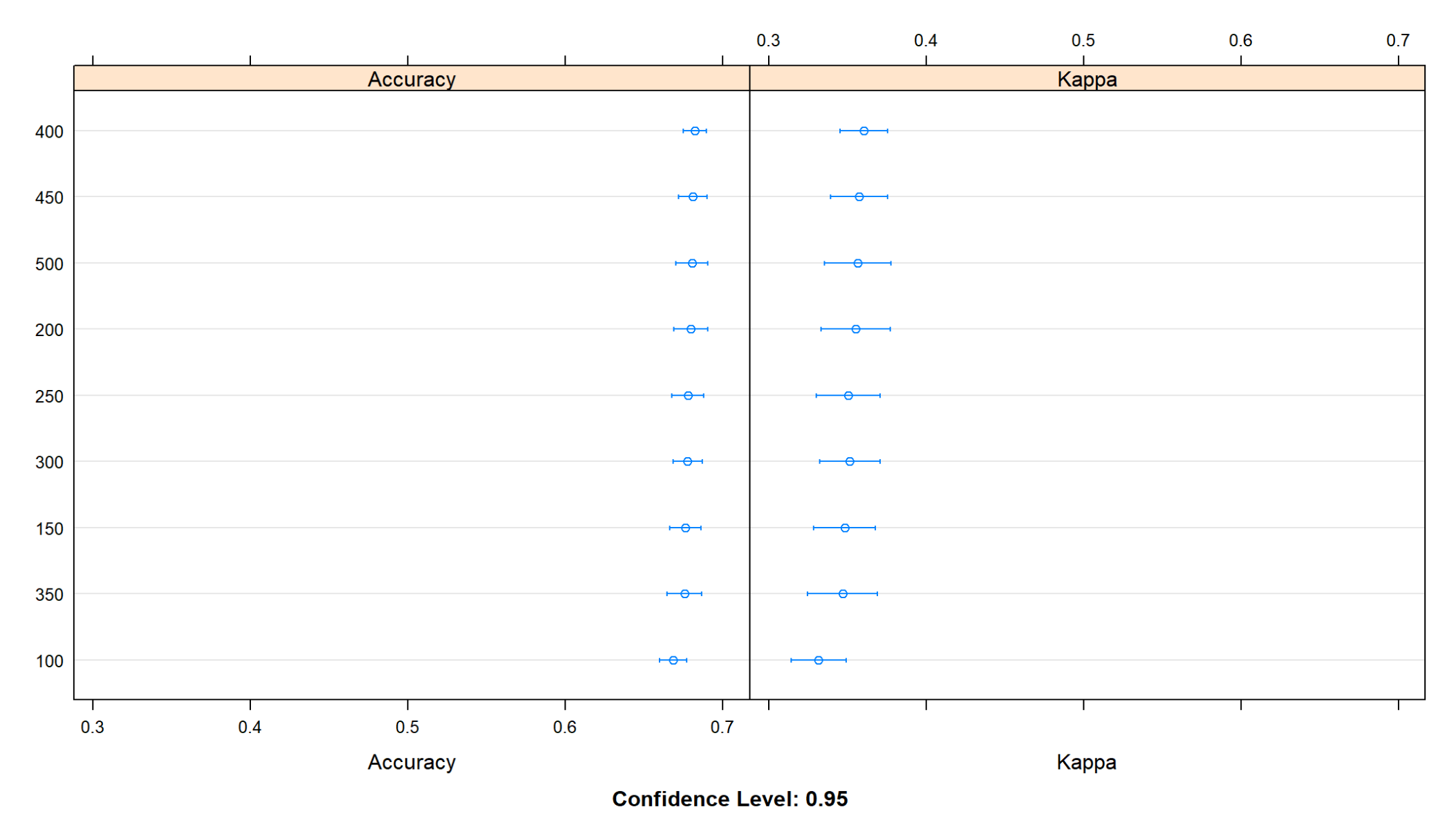
初步拟合发现（ntree=400,mtry=20），非平衡训练集并不能很好地学习小样本的特征。各类指标结果如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Unbalanced | | | | Balanced(Under Method) | | | |
| Accuracy | Recall 1 | Recall 2 | AUC | Accuracy | Recall 1 | Recall 2 | AUC |
| 0.951 | 0.033 | 0.999 | 0.516 | 0.706 | 0.623 | 0.710 | 0.666 |

为了进一步提高模型的性能，本文选择在平衡样本上的随机森林模型进行超参数ntree与mtry的3次10折网格搜索的交叉验证，即对集成树的数目以及每棵树使用的特征数目进行参数调整，搜索范围如下。

结果如下图所示。

从ntree参数随机森林包外误差准确率与Kappa值的平均值看，ntree=400是较好的参数，进一步查看ntree=400中，mtry的较好参数（mtry=25）。

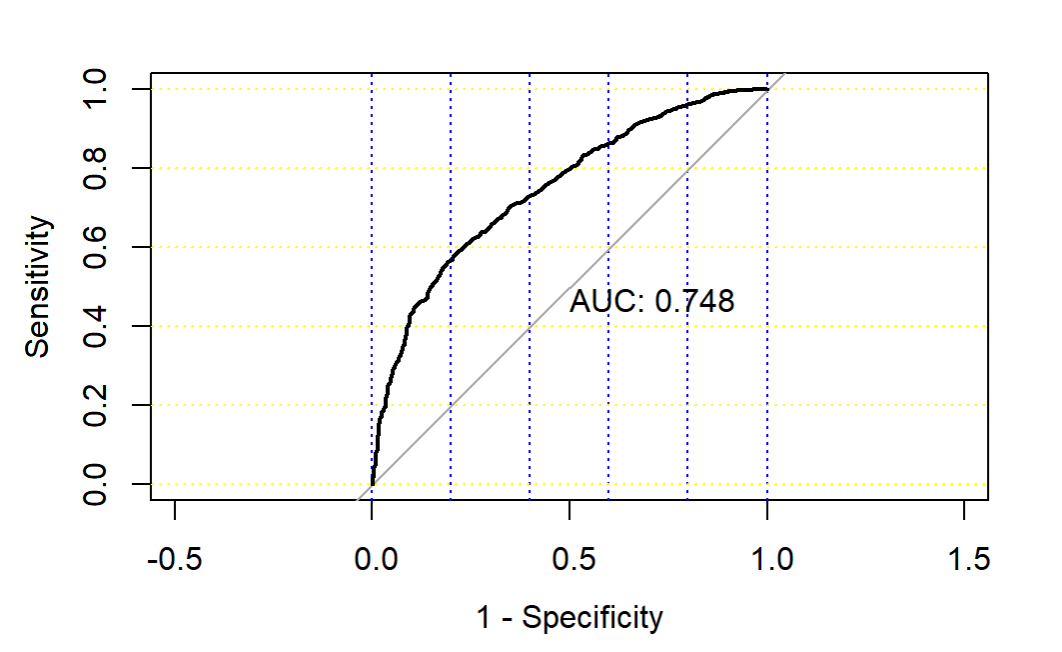


|  |  |  |
| --- | --- | --- |
| mtry | Accuracy | Kappa |
| 5 | 0.666649 | 0.327785 |
| 10 | 0.675075 | 0.345612 |
| 15 | 0.677619 | 0.35087 |
| 20 | 0.680611 | 0.356657 |
| **25** | **0.685933** | **0.367482** |
| 30 | 0.679616 | 0.354939 |

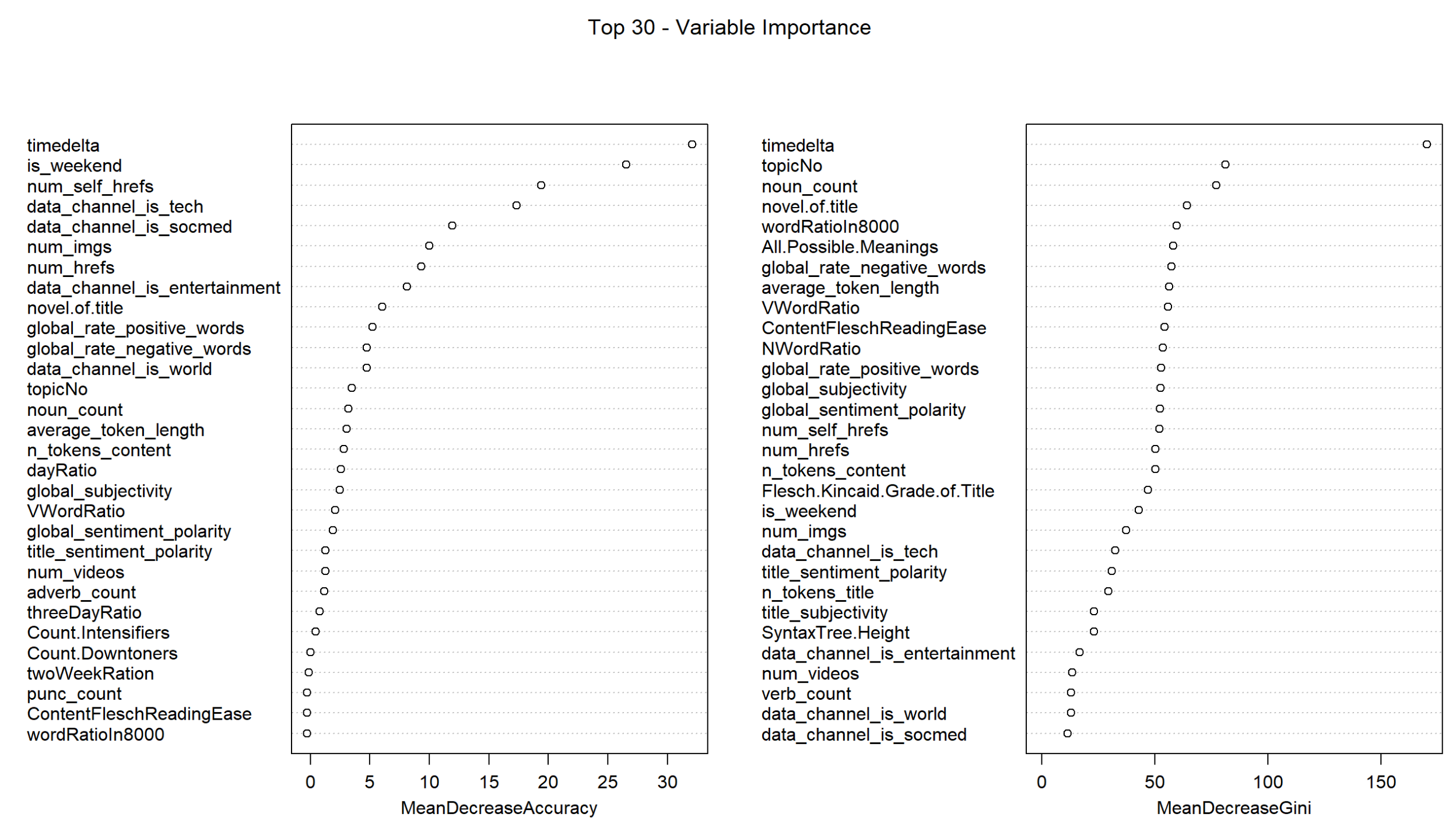
调参后指标变化如下：

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Recall 1 | Recall 2 | AUC |
| 0.705244 | 0.640411 | 0.708624 | **0.746** |

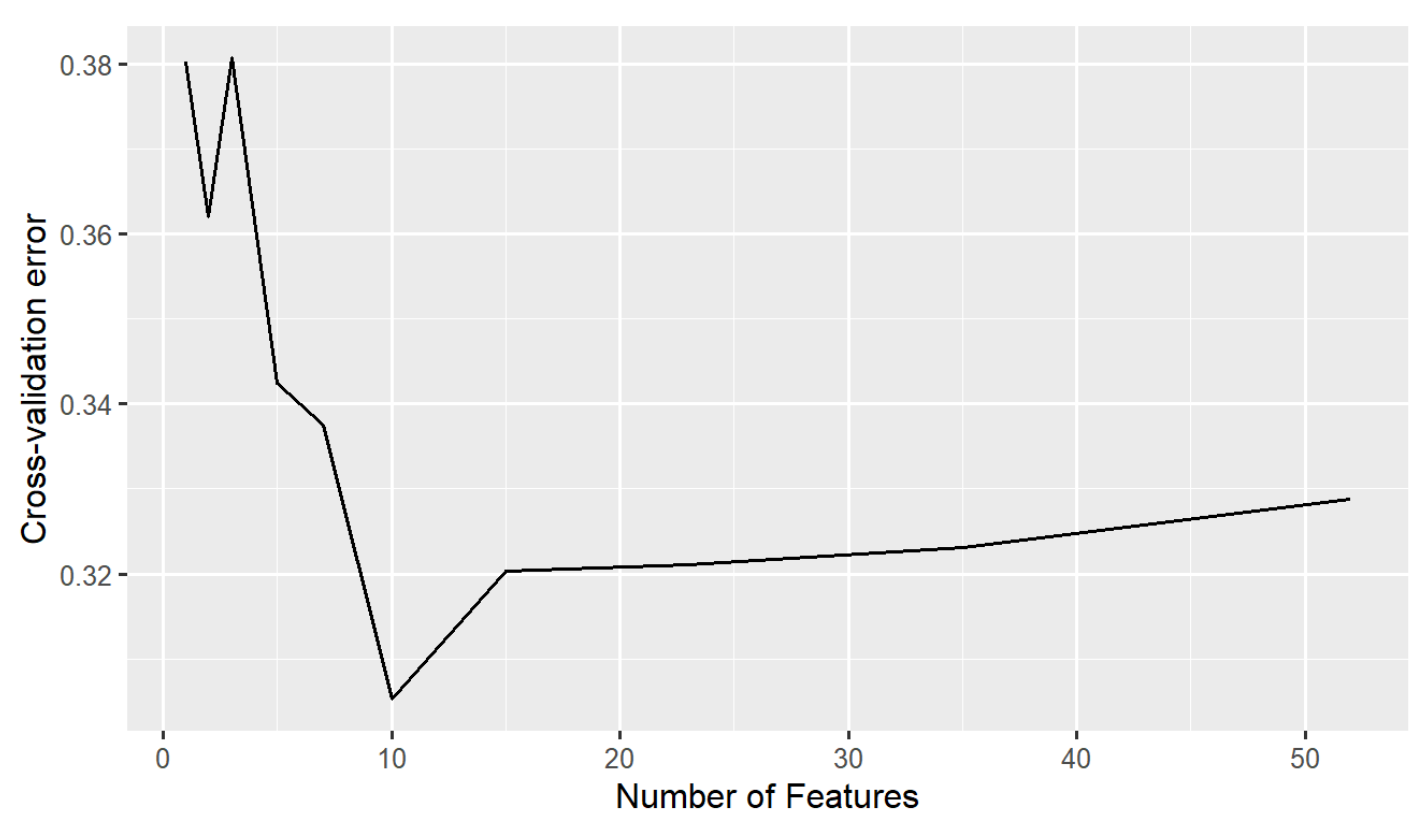
ROC曲线如下：



由于在数据预处理阶段并未做特征选择，为了模型与后续系统设计简洁性，同时兼顾效率，本文利用5次10折交叉验证以及随机森林模型中变量重要性（平均基尼下降值）来进行特征选择，变量重要性排序如下：

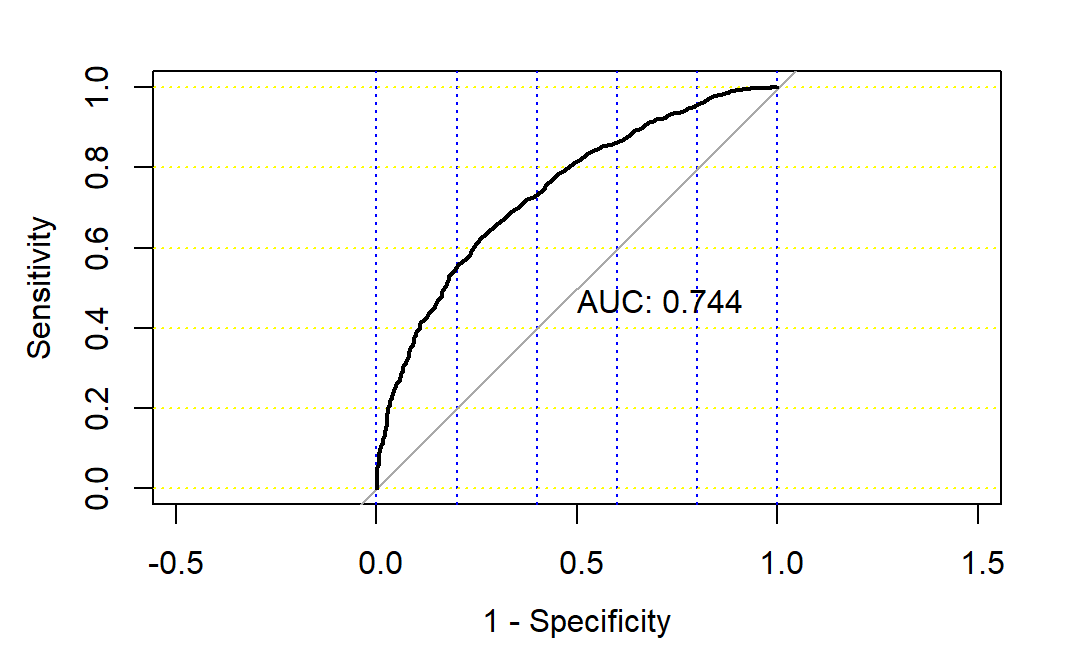


交叉验证结果如下：



然而，由于交叉验证所使用的指标为平衡样本集的包外误差，即平衡样本集boosting到的数据之外上所测得的准确率，因此交叉验证出的特征数目（Feature Num = 10）在不平衡测试集中并没有很好的效果。因此，经过手工验证，最终选定的特征数目为25。

特征选择后，模型ROC曲线如下：



最终特征如下：

|  |  |
| --- | --- |
| **Feature Name** | **MeanDecreaseGini** |
| timedelta | 1239 |
| noun\_count | 956 |
| average\_token\_length | 898 |
| topicNo | 875 |
| n\_tokens\_content | 859 |
| global\_subjectivity | 823 |
| wordRatioIn8000 | 757 |
| All.Possible.Meanings | 740 |
| novel.of.title | 738 |
| global\_rate\_positive\_words | 718 |
| VWordRatio | 712 |
| ContentFleschReadingEase | 700 |
| num\_herfs | 700 |
| NWordRatio | 688 |
| global\_sentiment\_polarity | 684 |
| global\_rate\_negative\_words | 686 |
| Flesch.Kincaid.Grade.of.Title | 606 |
| num\_imgs | 561 |
| num\_self\_hrefs | 472 |
| data\_channel\_is\_world | 449 |
| n\_tokens\_title | 434 |
| title\_sentiment\_polarity | 414 |
| data\_channel\_is\_entertainment | 405 |
| title\_subjectivity | 367 |
| is\_weekend | 358 |

##### 2.3.2 SVM优化

这里使用的是e1071包中的SVM函数，特征选择是由上面随机森林模型给出的较好的几个特征。由于SVM可以设置分类的权重，因此没有对样本中两个数量的类别进行平衡，而是在原数据集上训练。

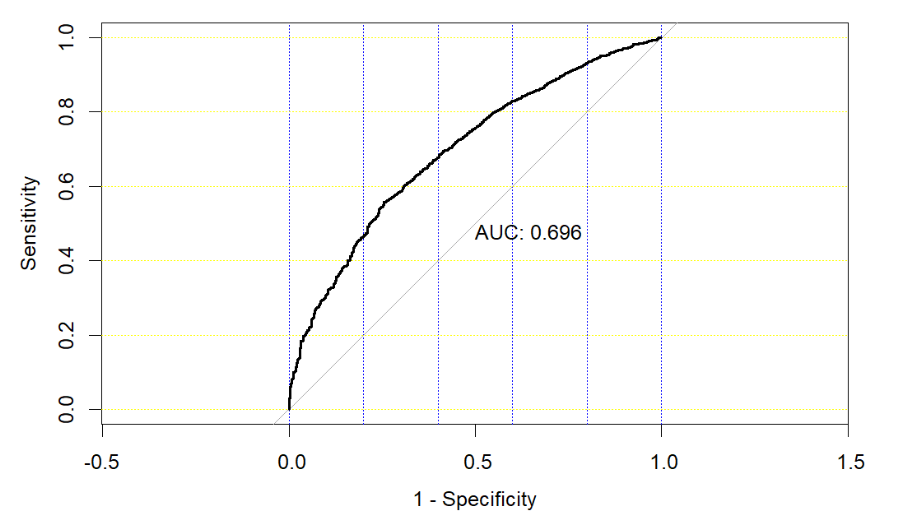
由于数据集是高维线性不可分的，因此选择高斯核作为核函数。svm函数的cost参数就是指Cost Function中的C参数，它是一个正则化参数，用来控制SVM的复杂度和容错性。C参数越大，表示对误分类的惩罚越大，SVM会尽量选择更少的支持向量，从而得到一个较窄的间隔和较复杂的决策边界；C参数越小，表示对误分类的惩罚越小，SVM会尽量选择更多的支持向量，从而得到一个较宽的间隔和较简单的决策边界。svm函数中的gamma参数用来控制核函数的形状和复杂度。gamma参数越大，表示核函数的影响范围越小，SVM会得到一个较复杂的决策边界；gamma参数越小，表示核函数的影响范围越大，SVM会得到一个较简单的决策边界。svm函数中的decay参数是指学习率衰减参数，用来控制SVM的学习速度和稳定性。decay参数越大，表示学习率衰减得越快，SVM会更快地收敛到一个局部最优解；decay参数越小，表示学习率衰减得越慢，SVM会更慢地收敛，但可能有更好的泛化能力。

考虑到本文的目的是发现较差的文章，使其改进，因此在应该注重在保证准确性的前提下提高较少类的召回率。

经过实践，发现在其他参数相同的情况下，使用全部特征的效果远不如使用部分特征，因此在这里我选用了上文随机森林选择出来的特征来预测分类。然后通过循环来寻找最好的参数的选择。在cost=2^(-5,5)，gamma=2^(-5,5)，decay=从(0.01,0.1)和五折交叉验证来寻找最优的参数组合，最后选择的参数如下如下：

|  |  |
| --- | --- |
| 参数 | 值 |
| kernel | "radial" |
| cost | 2^-5 |
| class.weights | c("1" = 1, "0" = 19) |
| gamma | 2^-4 |
| decay | 0.01 |

最终所得5%类的召回率为68.82%，准确性为60.65%，ROC曲线如下：



##### 2.3.3 神经网络优化

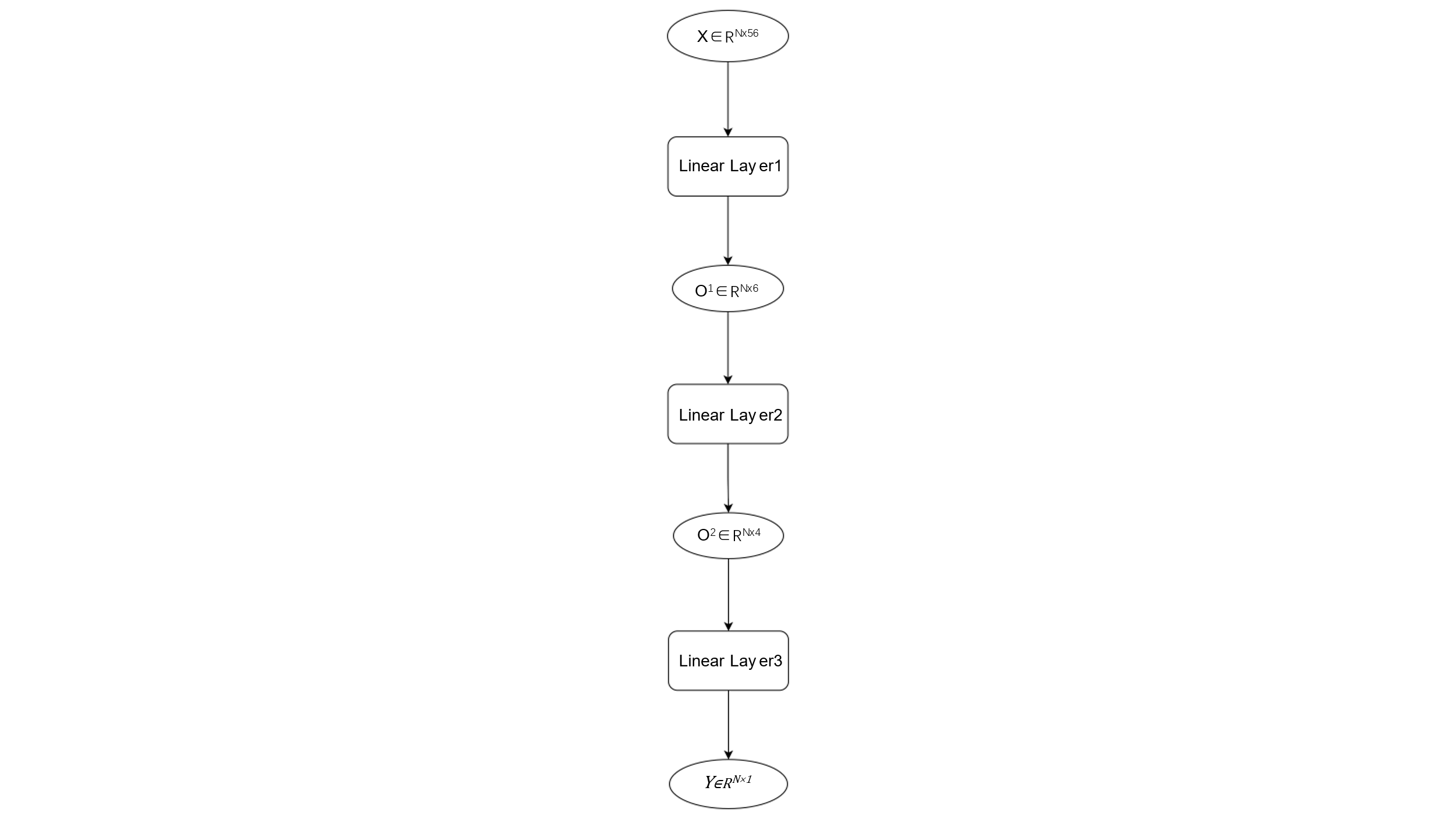
这里使用pytorch神经网络套路实现多维输入特征的二分类。大致可以分为六个步骤：1、设置随机数种子，实现pytorch结果的复现；2、划分训练集和测试集；3、构建模型，建立3个线性层，3个sigmoid非线性激活函数；4、选择损失函数；5、选择优化器；6、模型训练；7、测试及可视化展示。

在本模型中将数据集内所有数值型的属性（共56个）作为特征输入对shares属性进行二分类，并以9：1的比例将数据划分训练集与测试集。

调用torch.nn.Module模块构建神经网络模型。nn.linear(in\_features，out\_features，bias=True)用于设置神经网络中的一个线性层。其中in\_features为输入的神经元个数，out\_features为输出的神经元个数，bias用于控制偏置。

Linear就是对输入Xnxi进行了线性变换。即：Ynxo=XnxiWixo+b

其中W是模型需要学习的参数，W的维度为i\*o，b是o维的向量偏置，n为输入向量的行数，i为输入神经元的个数，o为输出神经元的个数。因本模型中共选用56个特征，故我们在构建模型时定义的三个模型参数分别为：W56x6，W6x4，W4x1。



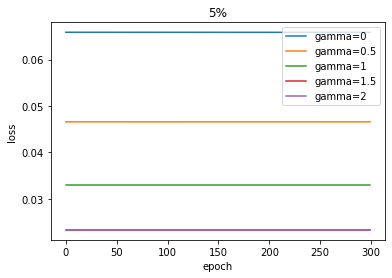
在线性层确定之后，利用激活函数sigmoid函数实现对（wx+b）的‘1-1’映射。完成对模型的构建。

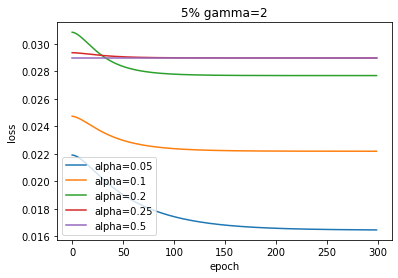
在机器学习模型中，每一个样本的预测值与真实值的差称为损失，损失函数是用来计算损失的函数，可度量一个模型进行每一次预测的好坏。损失函数主要用于模型的训练阶段。在每一个批次的训练数据送入模型后，通过前向传播输出预测值，然后损失函数会计算出预测值与真实值的差异值，即损失值。得到损失值后，模型通过反向传播去更新各个参数，来降低真实值与预测值之间的损失，使得模型生成的预测值往真实值靠拢，从而达到学习的目的。在训练完该模型后，此时模型通过反向传播后，已经使得每个参数都为最优。所以使用该模型进行预测得到的结果一定是接近真实结果的。分类任务中所使用的损失函数有：0-1loss、熵与交叉熵loss、softmax loss及其变种、KL散度、Hinge loss、Exponential loss、Logistic loss、Focal Loss等。

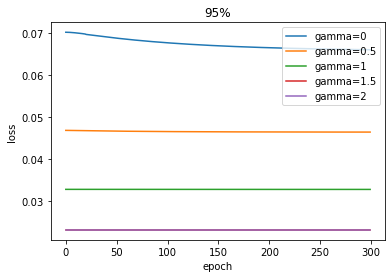
在本任务中，无论是以shares的5%为分位点进行二分，并记热门为“1”、正常为“0”，还是以shares的95%为分位点进行二分，并记正常为“1”，冷门为“0”，其正负样本分类均极不平衡。为了解决这个问题，在选择损失函数时我们选择构造Focal Loss函数。Focal Loss的具体公式如下：

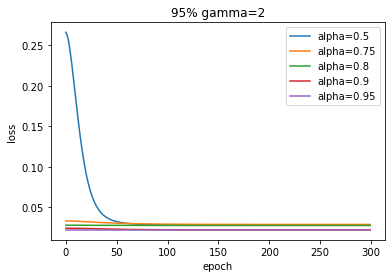
其中，超参数α代表了正样本的权重，可平衡正负样本之间的损失权重，如当正样本与负样本的比例为2：8时，α设置为0.8左右使效果最好；γ可用于控制简单/难区分样本数量失衡，且γ在Focal Loss中占主要地位；α与参数γ相互作用，随着γ的增大，α应该稍微降低。

Focal Loss参数设置：因在Focal Loss中γ占主要地位，故我们在构造Focal Loss函数时先对参数γ进行确定。通过查阅相关文献，我们选取了gamma=0、0.5、1、1.5、2五个数值进行测试，将数据集数据代入实例化，通过观察模型优化过程中的损失来选取最终参数。在选中γ后，再对超参数α进行抉择，选择alpha=0.05、0.1、0.2、0.25、0.5进行实例化测试。







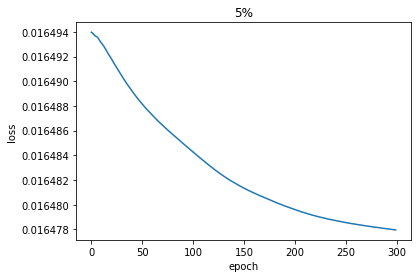


通过以上的可视化图可看出，在5%的二分类任务中，设置Focal Loss中参数γ为2，α为0.05时效果最好；而在95%的二分类任务中，设置Focal Loss中参数γ为2，α为0.95时效果最好。

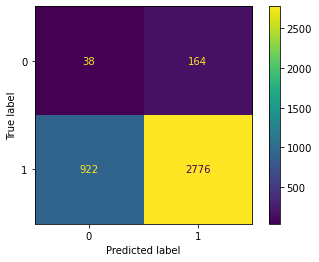
将模型和损失函数实例化后，需要对模型优化进行设置，我们控制模型的学习率为0.01。

在两次二分类任务中，将分别构建好的神经网络模型投入训练集中进行训练，且使其按照0.01的学习率自行循环优化三百次。分别绘制出两个模型各自训练及优化的结果。此后，将训练好的两个模型投入至对应的测试集中，绘制出相对应的混淆矩阵，并计算出精确率和召回率。

5%的分类任务中，模型训练及优化的结果如下：

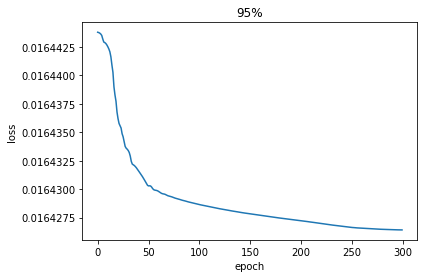


该模型在测试集中的表现如下图所示：

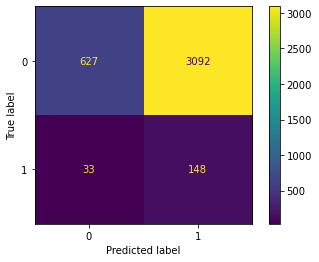


|  |  |  |  |
| --- | --- | --- | --- |
| Task | Precision\_score | Recall\_score1 | Recall\_score2 |
| 5% | 0.94421768 | 0.75067604 | 0.18811881 |

而在95%的分类任务中，模型训练及优化的结果如下所示：



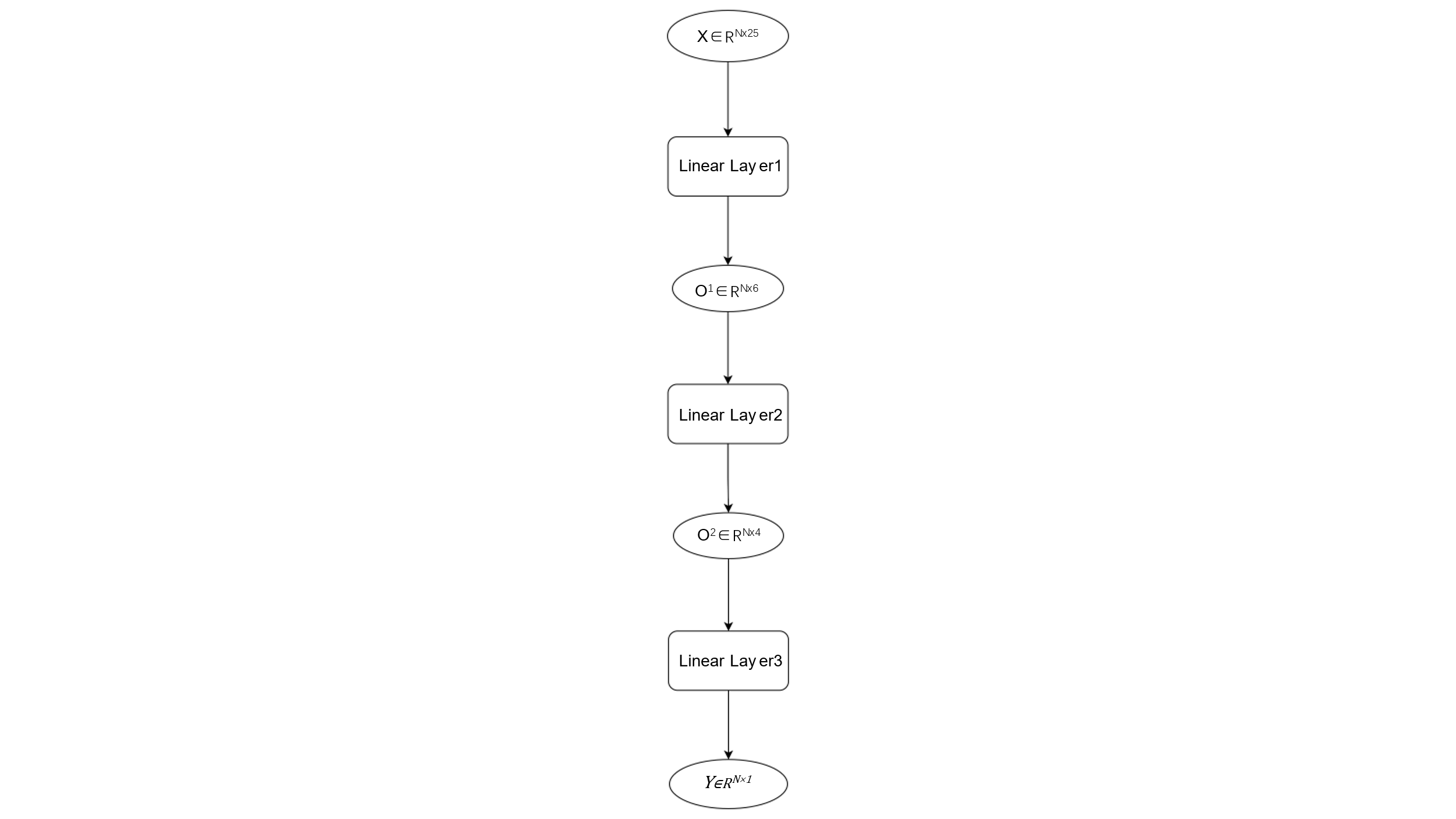
其在测试集中的表现如下图所示：



|  |  |  |  |
| --- | --- | --- | --- |
| Task | Precision\_score | Recall\_score1 | Recall\_score2 |
| 95% | 0.04567901 | 0.81767955 | 0.16859341 |

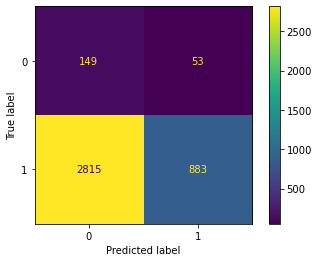
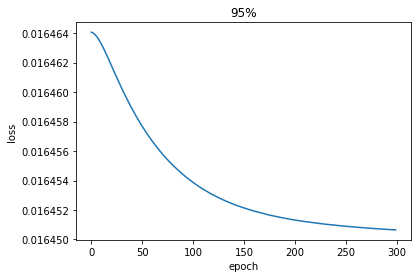
从模型投入测试集中测试的结果来看，两次分类任务最终的效果均不是很理想，即以原始数据集中56个数值型特征对shares进行二分类的效果并不好。在接下来，我们在两个任务中采用上文随机森林选择出来的25个数值型特征再对shares进行预测分类。

此时注意对神经网络模型的线性层进行调整，如下所示：



其余操作不改变，同样将模型投入训练集中训练优化，并在测试集中测试，绘制其混淆矩阵并计算出相应准确率与召回率。

在5%的分类任务中各项表现如下：



|  |  |  |  |
| --- | --- | --- | --- |
| Task | Precision\_score | Recall\_score1 | Recall\_score2 |
| 5% | 0.94337606 | 0.23877771 | 0.73762376 |

在95%的分类任务中各项表现如下：

## 模型解释与系统构想

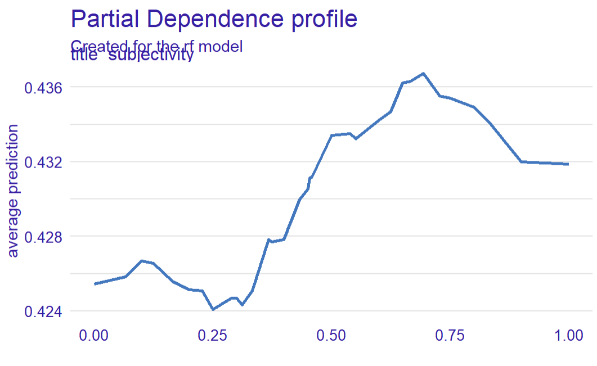
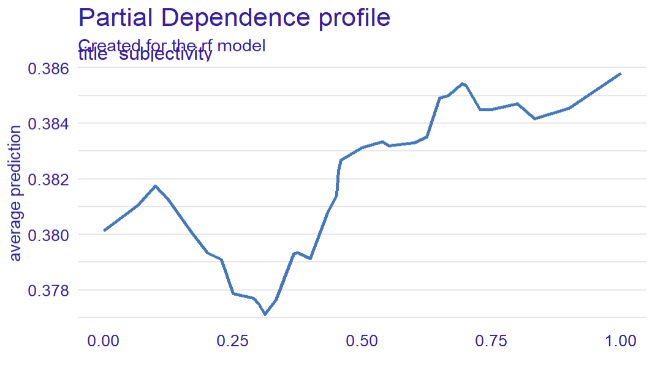
由于本文用于拟合的模型均不如树模型或逻辑回归能够直观解释特征如何影响结果，对后续供新闻工作者参考的决策系统构建会产生一定影响。

传统上，随机森林是通过组合弱分类器进行分类的学习模型。虽然随机森林在分类方面具有良好的表现，但其“黑盒子”特性使得模型的构建以及预测过程的细节对用户而言都是隐蔽的，对于我们面向的新闻工作者而言，并不能理解分类背后的原因。

因此，对于本文策划构建的系统来说，我们需要实现对于整体模型的变量影响解释以及对于单个实例的变量影响解释。为了实现这两个目标，本文将采用累积局部效应图（ALE）解释特征值的变化对预测值的影响，以及使用本地模型（LocalModel）的方法对单个实例的特征值影响进行探索。

### 基于累积局部效应图对于特征的解释

由于部份依赖图在该模型上的正负类影响难以看出差别（如下图，左为冷门类PDP，右为热门类PDP）。因此，采用能够消除一定特征相关性的累计局部效应图进行解释。



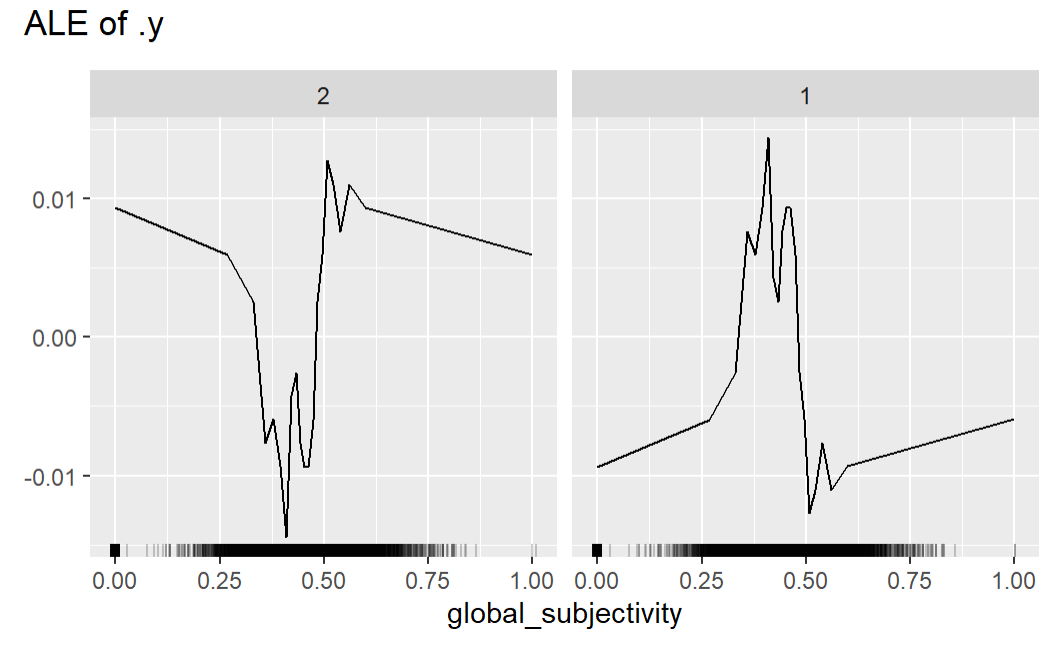
由累积局部效应的定义而言，其所衡量的即为排除其他特征影响后，被考察特征对目标函数值的边际贡献。累积局部效应图通过分箱、计算并累加局部效应值，可以用来展示特征对模型输出结果的边际贡献，着重刻画特征的非线性效应。

需要判断的特征较多，因此仅展示部分ALE图，其余放于附录之中。

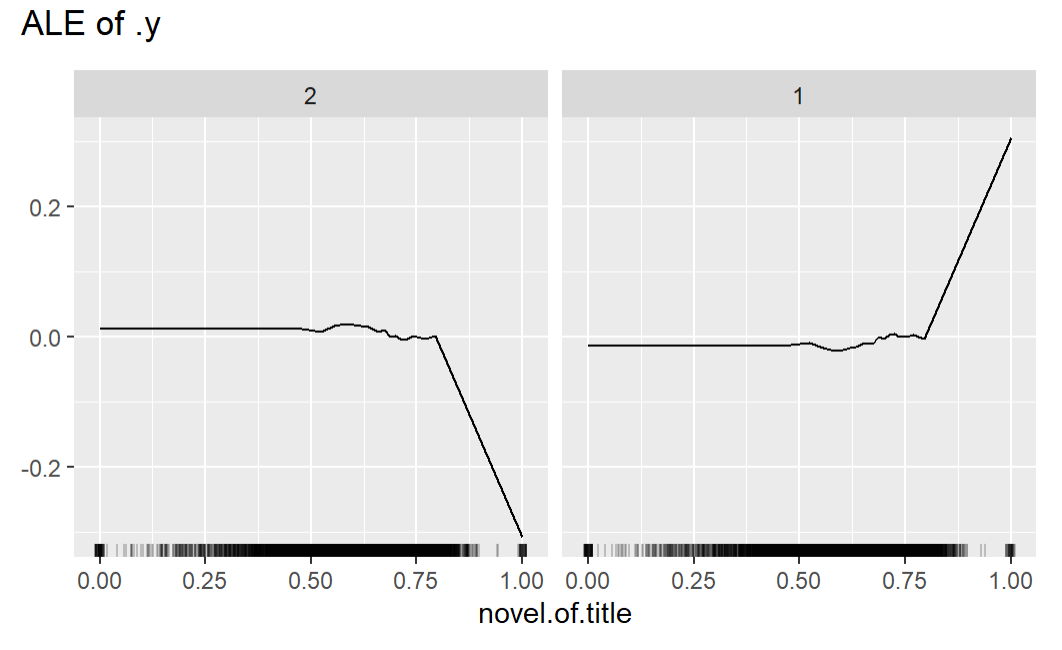
每个特征都有较为详细的语义，同时带有预先设想的对结果的影响，此处着重展示较难以解释影响的若干特征。

**文章主观程度**

由前文数据字典可得出，文章主观程度取值为0-1，值越大代表文章正文内容越主观。从图中可以看出，文章主观程度集中于0.3-0.5附近会极大增加文章的冷门概率。可以总结出，偏向客观事实输出和主观观点输出的文章能够更好地避免成为冷门文章，提高热度。

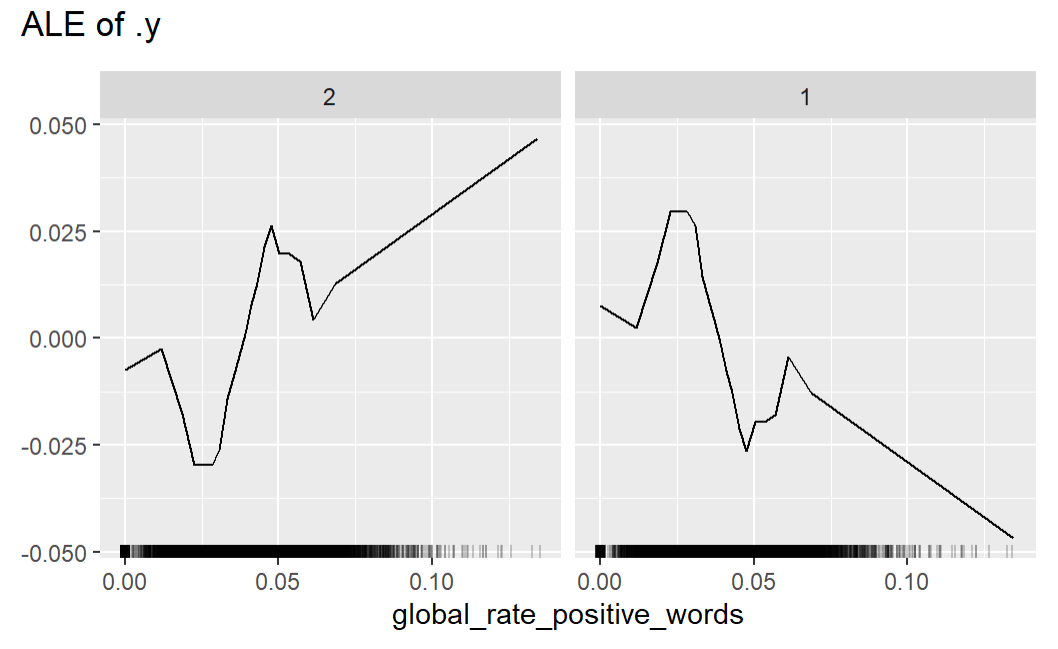
 **标题表述新奇性**

由前文数据字典可得出，标题新奇性的取值为0-1，值越大代表标题新奇程度越高。从图中可以看出，新奇性在小于0.8时，对结果几乎无影响；而大于0.8时，文章标题表述越新奇，文章成为冷门的概率越大。可以总结出，在Mashable这样一个博客文章平台上，标题越新奇，越与之前的文章标题不相似，文章成为冷门的概率越大。



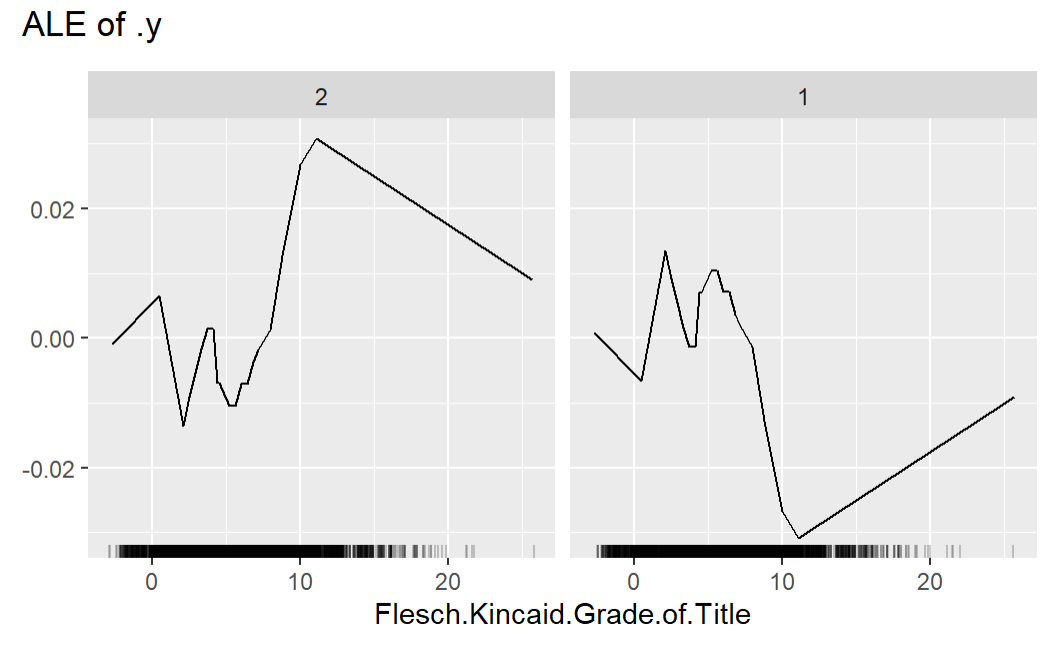
**积极词的占比**

由前文数据字典可得出，标题新奇性的取值为0-1，值越大代表正文积极词占比越高。从图中可以看出，积极词占比对文章热度的影响呈严重非线性关系，难以从现实解释趋势变化的意义。但从总体趋势上来看，积极词占比越低，文章成为冷门文章的概率越高。



**标题可读性指数**

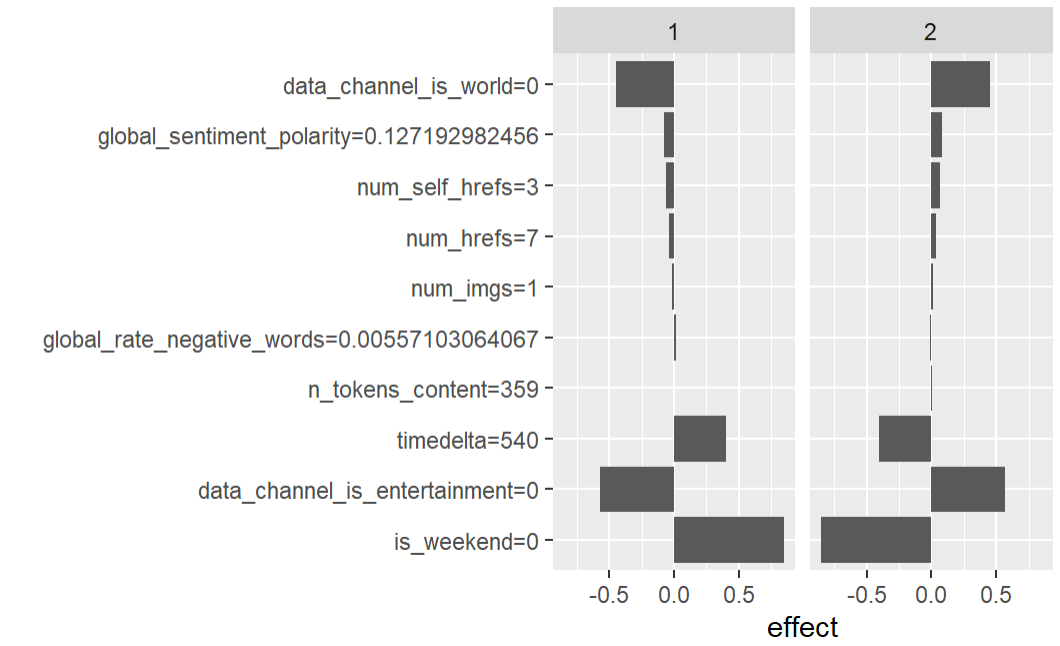
由前文数据字典可得出，标题可读性指数值越大代表标题越难以读懂。此特征趋势波动也较大，但是可根据总体趋势划分为三个部分：小于5，5-11，11以上。本文猜测，由于读者群体类别不同，导致类别的文章具有最适应的难度级别，从而能够吸引目标读者群体。



### 基于Local Model对于单个实例特征影响度量

最后系统需要针对单个新闻工作者给出合理建议，因此需要针对单个实例给出优化建议，即对局部实例进行解释性研究。前文给出累积局部效应图，可依据对累积局部效应的解释以及单个实例特征影响的结果对实例进行优化。

本文使用iml包中的LocalModel对单个实例进行变量影响研究，下图为实例结果。



## 总结

### 成果

本文结合过往研究，提出了在在线文章冷启动的情况下识别冷门文章的任务，并且针对模型的特征对结果进行解释，向内容生成者提供相关建议。

在特征构建的阶段中，本文使用NLP技术和机器学习方法，根据过往研究出的一些特征与本文构建的体系进行人工构建。

在模型拟合阶段，本文使用随机森林、支持向量机与神经网络三个模型进行拟合，并且分别针对不平衡的样本采用了采样、调整SVM模型参数与修改神经网络损失函数方法提高模型性能，通过调参过程得到了三个最优模型。最好的结果由随机森林实现，ROC曲线下的总面积为0.748，这相当于一个可接受的区间，比过往相同数据集均衡二分任务提高了一些性能。此外，本文还基于平均基尼下降值（MeanDecreaseGini）分析了随机森林特征输入的重要性，在平衡特征数目和模型性能的情况下筛选出了25个变量，供后续模型解释使用。

最后，在模型解释阶段，根据现实意义与累积局部效应图解释了随机森林中特征输入的影响以及变化趋势，根据Local Model输出单个实例的特征影响，为后续内容生成者的建议系统提供了一定方法和基础。

### 不足

在特征构建阶段，由于本文三位作者欠缺系统的自然语言处理基础，因此对于内容与语义特征的构建较为粗糙；同时由于维基百科未收录2015年前词条的热度信息，同时谷歌热度是基于当前词条的热度相对值，因此热度信息未能纳入特征中进行模型拟合。

在模型解释阶段，由于本文三位作者欠缺系统的机器学习与深度学习可解释性理论的基础，因此对于模型的解释较为粗糙，如没有PDP与ICE能体现出的特征异质效应，同时相关性检验结果并未采纳，可能会导致ACE图依然会由于特征之间的相关性产生一定差错。同时由于模型解释方法并不稳健，ACE的间隔数设置可能会导致最终结果的不同。

对于实例模型解释来说，实际上有更加通用的方法，能够深入实例的理解。

## 参考文献

Arapakis, I., Cambazoglu, B. B., & Lalmas, M. (2014). On the Feasibility of Predicting News Popularity at Cold Start. *Social Informatics, Socinfo 2014, 8851*, 290-299. Retrieved from <Go to ISI>://WOS:000354775400021

Liao, D. L., Xu, J., Li, G. F., Huang, W. J., Liu, W. Q., & Li, J. (2019). Popularity Prediction on Online Articles with Deep Fusion of Temporal Process and Content Features. *Thirty-Third Aaai Conference on Artificial Intelligence / Thirty-First Innovative Applications of Artificial Intelligence Conference / Ninth Aaai Symposium on Educational Advances in Artificial Intelligence*, 200-207. Retrieved from <Go to ISI>://WOS:000485292600025

Piotrkowicz, A., Dimitrova, V., Otterbacher, J., & Markert, K. (2017). *Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook*.

李泽月. (2019). *基于新浪新闻数据分析的新闻热度预测方法研究.* (硕士). 中国地质大学(北京), Available from Cnki

## 附录

### 特征构建

请打开preprocess.ipynb文件查看详细信息。

### 数据字典

详情请见Structure.md

时间特征

timedelta： 距离收集时间的间隔，以天为单位。

Date： 收集数据的日期，格式为年-月-日。

weekday\_is\_monday： 是否是星期一，是为1，否为0。

weekday\_is\_tuesday： 是否是星期二，是为1，否为0。

weekday\_is\_wednesday： 是否是星期三，是为1，否为0。

weekday\_is\_thursday： 是否是星期四，是为1，否为0。

weekday\_is\_friday： 是否是星期五，是为1，否为0。

weekday\_is\_saturday： 是否是星期六，是为1，否为0。

weekday\_is\_sunday： 是否是星期日，是为1，否为0。

is\_weekend： 是否是周末，是为1，否为0。

isHoliday： 是否是节假日，是为1，否为0。

HolidayName： 如果是节假日，节假日名称（Not Holiday表示不是节假日）。

HolidayDay： 如果是节假日，是节假日的第几天（从1开始计数）。

HolidayDaysLeft： 如果是节假日，距离节假日结束还有几天（最后一天算1）。

dayRatio： 文章发表的前一天相较于前两天的热度增长率增长

threeDayRatio： 文章发表的前三天相较于前六到三天的热度增长率增长

weekRatio：文章发表的前一周相较于前两周的热度增长率增长

twoWeekRatio：文章发表的前两周相较于前四到二的热度增长率增长

栏目特征

data\_channel\_is\_lifestyle： 是否属于生活方式栏目，是为1，否为0。

data\_channel\_is\_entertainment： 是否属于娱乐栏目，是为1，否为0。

data\_channel\_is\_bus： 是否属于商业栏目，是为1，否为0。

data\_channel\_is\_socmed： 是否属于社交媒体栏目，是为1，否为0。

data\_channel\_is\_tech： 是否属于科技栏目，是为1，否为0。

data\_channel\_is\_world： 是否属于世界栏目，是为1，否为0。

### 热门文章分类

根据采样过程选取最好的过采样，随机森林模型最终混淆矩阵如下：



### 所有连续变量ALE图

