1.1数据前处理

1.1.1法院分类

|  |
| --- |
| 北京市高级人民法院 |
| 北京市第一中级人民法院 |
| 北京市高级人民法院 |
| 北京市房山区人民法院 |
| 山东省淄博市中级人民法院 |
| 北京市第三中级人民法院 |
| 北京市高级人民法院 |
| 北京市高级人民法院 |
| 北京市高级人民法院 |
| 北京市高级人民法院 |
| 北京市高级人民法院 |
| 北京市第二中级人民法院 |
| 江苏省南通市中级人民法院 |
| 上海市高级人民法院 |

此为部分案例判决法院实例，特征构成中，按照

1. 中级人民法院；
2. 高级人民法院；
3. 最高人民法院
4. 地区性人民法院
5. 其他法院

来进行分类

1.1.2原告分类

|  |
| --- |
| 马则乃 |
| 信阳市南湾湖风景区南湾街道办事处郑家冲九组 |
|  |
| 孙友 |
| 陈建元、周和余、余国雄、余秀宝 |
| 张静静 |
| 顾立清 |
| 刘冬生、王文斌 |
| 广东省韶关市人民政府 |
| 无锡市辉宏企业管理咨询有限公司 |
| 蔡孝敏、杨秀婷 |
| 张利华 |
| 谭楚云 |
| 牛利敏 |

此为部分案例原告实例，特征构成中，按照

1. 个体
2. 法人
3. 集体

来进行分类

1.1.3被告分类

|  |
| --- |
| 北京市怀柔区人民政府 |
| 北京市公安局 |
| 北京市通信管理局、中华人民共和国工业和信息化部 |
| 北京市公安局东城分局、北京市公安局 |
| 如皋市综合行政执法局、南通市城市管理综合行政执法局、如皋市民生住房开发有限公司 |
| 上海市徐汇区住房保障和房屋管理局、上海市住房和城乡建设管理委员会 |
| 上海市监狱管理局、上海市人民政府 |
| 天津市住房和城乡建设委员会、中华人民共和国住房和城乡建设部 |
| 民和回族土族自治县住房和规划建设局、海东市住房和规划建设局 |
| 信阳市人民政府 |
|  |
| 北京市农业农村局、中华人民共和国农业农村部 |
| 上海市闵行区人民政府、上海市人民政府 |
| 北京市公安局大兴分局、北京市大兴区人民政府 |
| 北京市农业农村局、中华人民共和国农业农村部 |
| 北京市朝阳区农村合作经济经营管理站、北京市朝阳区人民政府 |
| 秀水村民委员会大朱家村民小组 |
| 无锡市司法局、江苏省司法厅 |
| 中华人民共和国公安部 |
| 湘潭市公安局岳塘分局、湘潭市人民政府 |
| 湘潭市公安局岳塘分局、湘潭市人民政府 |
|  | |

此为部分案例被告实例，特征构成中，按照

1. 人民政府
2. 公安局
3. 税务局

国家部门（例如中华人民共和国农业农村部）

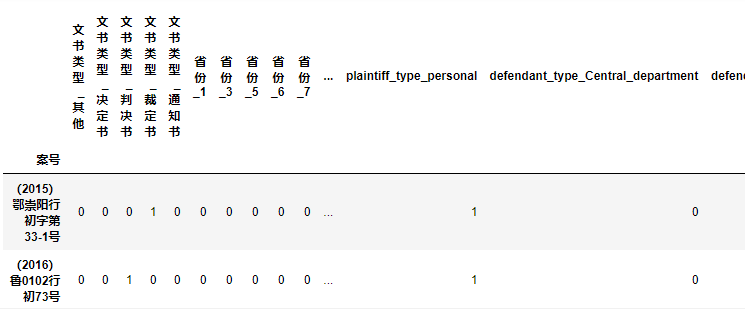
1. 其他局类部门（例如环保局，质监局）
2. 街道办事处
3. 各级委员会
4. 人民政府与公安局
5. 人民政府与其他
6. 其余未分类

来进行分类

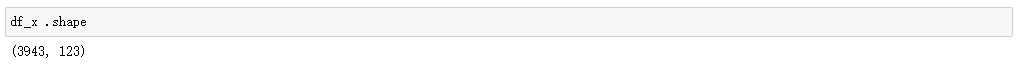
1.1.4TF-IDF高频词

1.2特征构成

将所有分类项转化为0-1变量，如下图所示



总计有123个特征，均为0-1变量



训练集为败诉胜诉的0-1变量



1.3特征学习

利用5种分类算法进行集成，利用KFold来避免标签泄露。

1.3.1算法介绍

xgboost，lightgbm，svr，岭回归，随机森林

1.3.2训练模型

通过交叉训练，观察各个分类算法对于败诉胜诉分类的损失函数的多寡，结果如下：

|  |  |
| --- | --- |
| xgboost | 0.3534 |
| lightgbm | 0.3564 |
| svr | 0.4577 |
| 岭回归 | 0.3672 |
| 随机森林 | 0.3541 |

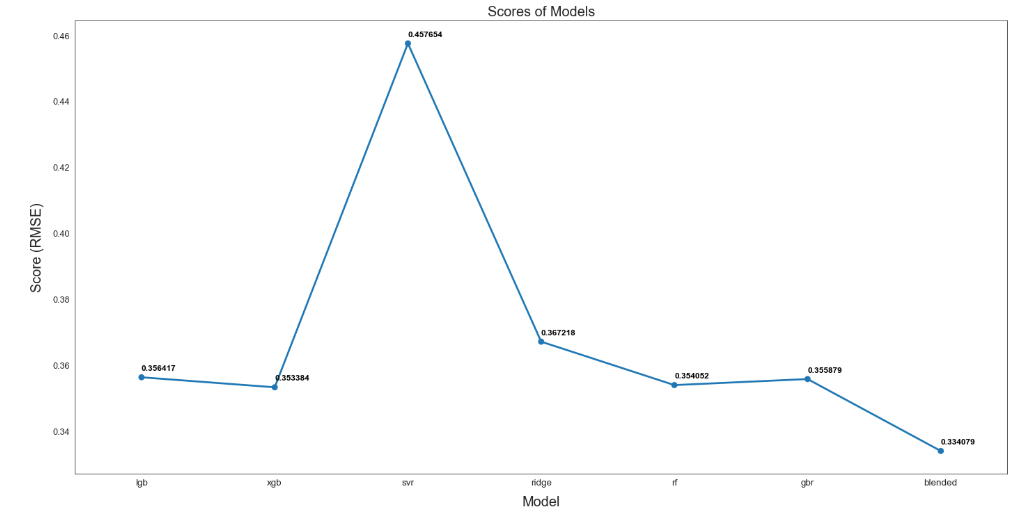
根据其损失函数对综合学习器进行加权，得到一个强学习器，此强学习器对于各个算法进行stacking集成后线性回归。

最后我们得到6个分类器，5个基础的分类器和一个集成的强分类器，再利用损失函数加权形成最终的第七个分类器，这是为了使得学习器更加稳健，不至于过拟合：

权重如下

|  |  |
| --- | --- |
| xgboost | 0.1 |
| lightgbm | 0.1 |
| svr | 0.2 |
| 岭回归 | 0.1 |
| 随机森林 | 0.05 |
| 强学习器 | 0.35 |

最终得到的综合学习器，损失函数值为0.33407，比任何一个弱学习器都要小



1.4预测和拟合

1.4.1预测

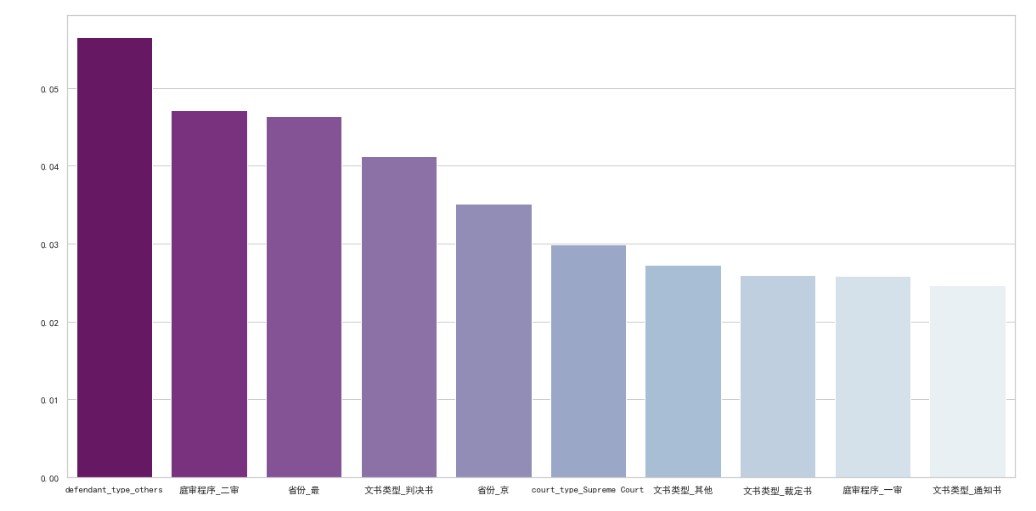
因为使用的是线性回归模型，在测试AUC和ACC时将预测值-0.5后，大于0得标记为1，小于0的标记为0。

准确率为84.32%

AUC值为66.03%

1.4.2评估各个变量对于模型的影响

此处以xgboost为例，观察模型中各个特征值的重要性



由图可见，模型中最为重要的参数是

1. 被告是否为“其他类型”
2. 庭审程序是否二审',
3. '省份是否分类为“最高人民法院”
4. 文书类型是否判决书',
5. '省份是否分类为北京,
6. 法院类型是否为“最高人民法院”
7. 文书类型是否为“其他”

'（8） '文书类型是否为“裁定书”

（9）'庭审程序是否为一审

（10）文书类型是否为“通知书”

Top20的因素及其重要性大小为

defendant\_type\_others 0.056482

庭审程序\_二审 0.047097

省份\_最 0.046405

文书类型\_判决书 0.041229

省份\_京 0.035096

court\_type\_Supreme Court 0.029846

文书类型\_其他 0.027301

文书类型\_裁定书 0.025965

庭审程序\_一审 0.025830

文书类型\_通知书 0.024640

court\_type\_high 0.022351

defendant\_type\_other\_bru 0.017598

庭审程序\_其他 0.014235

defendant\_type\_Central\_department 0.013651

plaintiff\_type\_legal entity 0.013541

庭审程序\_再审 0.013217

省份\_平 0.012884

省份\_赣 0.012867

省份\_琼 0.012767

省份\_陕 0.012700