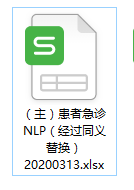
**关于候诊时间预测的初步阶段总结**

1. **数据样式**

源数据：

数据特征内容：['His流水号', '急诊系统流水号', '门诊号', '就诊时间', '年龄', '性别', '分诊科室', '分诊等级', '分诊去向', '分诊时间','抢救上床时间', '体温类型', '体温', '脉搏', '收缩压', '舒张压', '呼吸', 'SpO2', '微量血糖', '疼痛评分', 'GCS评分', 'PHI评分', '来诊主诉', '现病史', '既病史', '药物过敏史', '诊断', '处理意见', '注意事项', '时长', '没有时长的主诉', '主诉分词', '编码']，数据量共350626条，涵盖了2016年、2017年、2018年的数据

根据需要，从 '分诊时间' 特征中进行筛选出2018年的数据，数据量共有12万8645条。

1. **数据处理**

针对18年的数据，进行特征的选择与处理

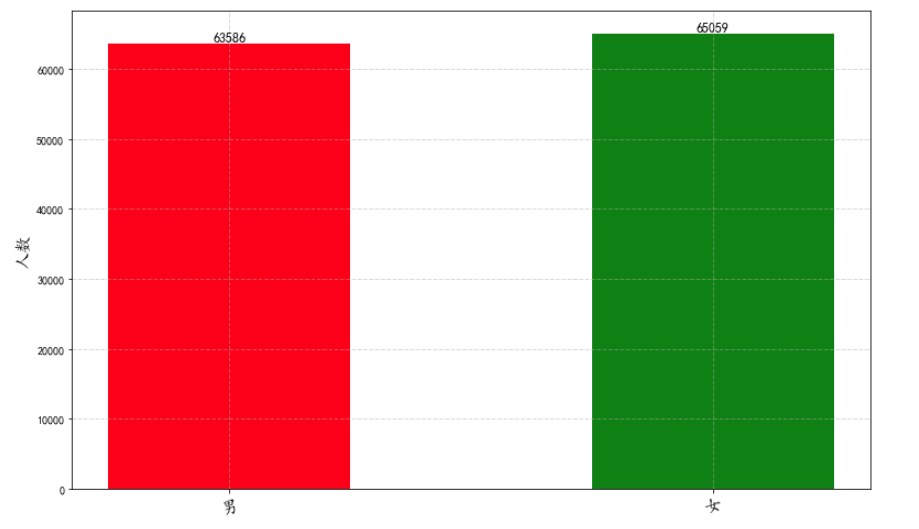
（1）、数据中【'His流水号', '急诊系统流水号', '门诊号'】这部分数据是属于医院相关的，专属病人的代号，在数据模型训练中利用不上，做去除处理。

（2）、目标：候诊时间，为等候的时长，是模型训练的target，是由【'就诊时间'】减去【‘分诊时间’】得出，时间单位为分钟。

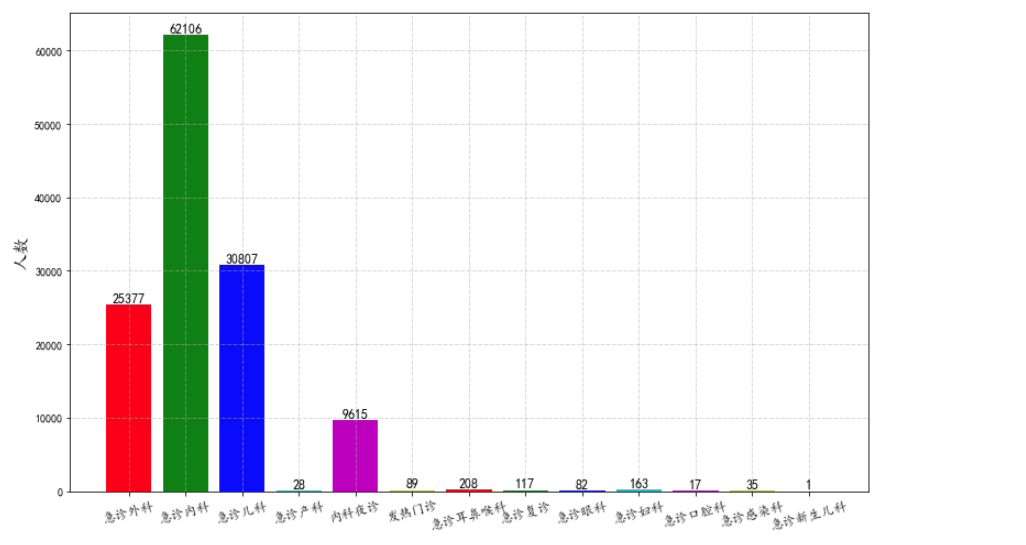
（3）、数据中特征【‘年龄’】的样式是字符串的类型，有数值的同时也标注了年(‘Y’)和月(‘M’) 的区别。在处理中，有‘Y’的标志直接取其数值作为年龄，对于‘M’标志的取其数值后除以12，换算成以年为单位的数值。

（4）、数据中特征【‘性别’】，按照惯例，采用0/1来代替，‘男’为0，‘女’为1。

性别人数的分布情况：

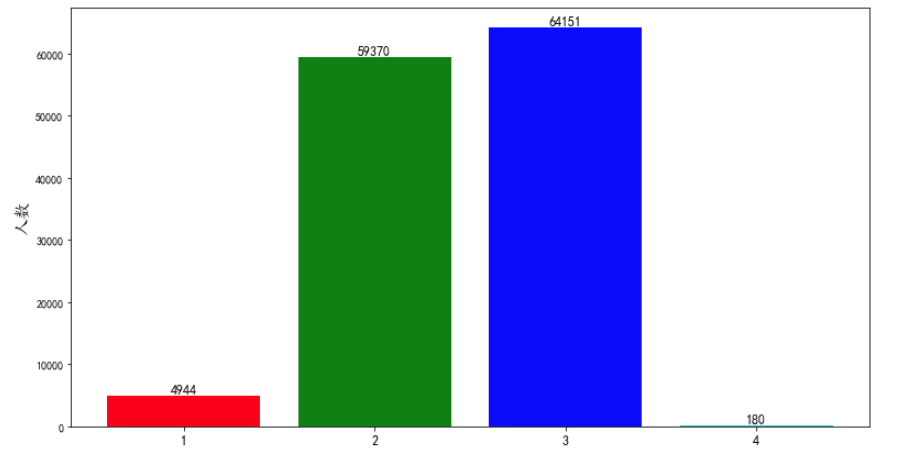


（5）、特征【‘分诊科室’】，含有的数据为['急诊外科', '急诊内科', '急诊儿科', '急诊产科', '内科夜诊', '发热门诊', '急诊耳鼻喉科', '急诊复诊', '急诊眼科', '急诊妇科', '急诊口腔科', '急诊感染科', '急诊新生儿科']13种数据样式，分别以1到13的数值进行代替。



（6）、特征【‘分诊等级’】的数据样式均为数值型【1,2,3,4】

其人数分布为：



1. 、特征【‘分诊去向’】，数据样式仅为中文样式【‘医疗室’，‘抢救室’】，‘医疗室’的数量占比最多，有12万5195条，‘抢救室’有3450条数据。由于是中文样式，转换数字目前分别用【1,2】来进行代替。
2. 、特征【‘分诊时间’】不作特征选项，但从该特征进行分裂，形成新的特征【‘分诊-月’，‘分诊-日’，‘分诊-时’，‘分诊-分’】四个新特征进行如模型训练的。这里没有‘分诊-年’，这是一个过去的数值，新数据后来的不会再形成循坏的数值，没有‘分诊-秒’是因为‘分诊时间’中的数值中，‘秒’的部分全部都是0。
3. 、特征【'抢救上床时间'】为穿越数据，不作特征考虑。
4. 、特征【'体温类型'】的数据样式为中文样式的['耳温', '腋温']，数据中只有7条数据是描述为‘腋温’，由于是中文样式，转换数字目前分别用【1,2】来进行代替。
5. 、特征【‘体温’】数据中，数据样式中除了有整型的数据外，还有其他不符合的字符存在，可能是技术人员录入时出错导致的。在进行修改时，按照合理的原则进行修正，例如：

这方面的数据样式为带有其他不符合的字符存在时，但却能够直观的看到起数值的大小，故按照合理的规则进行修正。除此之外对于一些空值，和一些异常值的存在，统一修正为37。

1. 、特征【'脉搏'】对空值的进行填充平均值89.96（平均值根

据所有‘脉搏’的存在的数值和除以所有的人数）

1. 、特征【'收缩压'】对空值的进行填充平均值138.404（平

均值根据所有‘收缩压’的存在的数值和除以所有的人数）

1. 、特征【'舒张压'】对空值的进行填充平均值81.03（平

均值根据所有‘'舒张压’的存在的数值和除以所有的人数）

1. 、特征【'呼吸'】对空值的进行填充平均值19.3（平

均值根据所有‘呼吸’的存在的数值和所有的人数）

（16）、特征【'SpO2'】，数据样式中除了有整型的数据外，还有其他不符合的字符存在，对其进行合理性原则修正，对于空值或者没有值的数据补充为0。

（17）、特征【‘疼痛评分’】都为一些不合理、且无法理解的数据存在，因此这个特征，不作模型训练特征。

（18）、特征【'GCS评分'】对空值的进行填充平均值15（平

均值根据所有‘呼吸’的存在的数值和所有的人数）

1. 、特征【'PHI评分'】对空值的进行填充0。
2. 、特征【‘来诊主诉’】参照‘分级’的项目模型的做法，

对主诉中的词语进行提取，按照项目中的统一编码来进行数值代替，形成新特征【‘编码’】。

1. 、特征【'现病史', '既病史', '药物过敏史', '诊断', '处理意见', '

注意事项', '时长', '没有时长的主诉'】这方面的数据样式均为中文句子表示，大多数为空值。由于其中间或有穿越数据，且中文句子中需要进行词语提取、词语替换、词语编码等NLP（自然语言处理）的过程，目前不考虑作为特征进行训练模型。

**综上所述**，经过数据分析与处理后，符合进行模型训练的特征为：['年龄', '性别', '分诊科室', '分诊等级', '分诊去向', '体温类型', '体温', '脉搏', '收缩压', '舒张压', '呼吸', 'SpO2', '微量血糖', 'GCS评分', 'PHI评分', '分诊-月', '分诊-日', '分诊-时', '分诊-分', '脉压差']；模型训练的target为：【‘候诊/分钟’】。

1. 模型训练

采用上述中数据处理方法形成的特征数据进行模型搭建，目前已初步建立的过往常用的模型lightGBM模型，模型参数采用默认的参数，往后将根据实际情况进行参数调整，或者选择神经网络的模型。

把数据进行按照8:2 的比例进行随机分割训练集和测试集，训练集的数量是10万2916，测试集是2万5729条

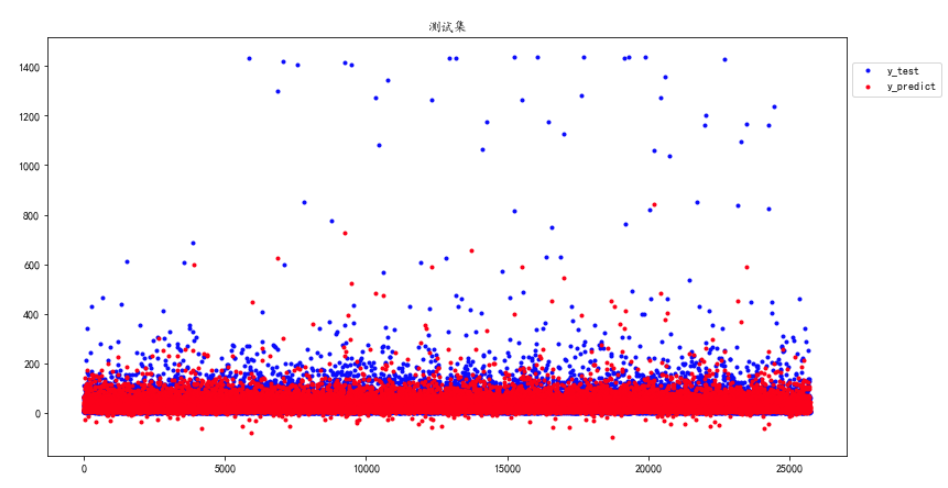
1. 、模型结果

目前初步的探索，在构建一个新特征【‘脉压差’】，由数据中的【‘收缩压’-‘舒张压’】获得。

目前最优的模型的分数是：

MAE（平均绝对误差）：23.5956（分钟）

R方决定系数：0.28（越接近1越好）



从上图来看，预测出来的红点的值基本覆盖在了真实值蓝点的密集区域，但对于真实值在200周围及以上的数据，拟合得并不好。

目前定义一个标准是，预测值最好不超过真实值的15分钟。在汇总真实值与预测值的数据进行统计，符合标准的数据量在8445条，只占了测试集数据量的约33%。但如果预测值小于真实值的（即预测值减去真实值的差的绝对值在15以内）来看，数据量达到1万2665条，占测试集的数据量约50%。如果定义的误差是30 分钟，则符合标准的数据量有1万9197条，占测试集数据量的约75%。因此存在有一定的模型优化空间。

1. 建议

目前的病人候诊时间的初步模型的评分在23.6（MAE），这方面的数据进行训练时来源于病人本身，即通过病人的自述主诉以及生化指标的数据呈现；但候诊时间的本身对于医疗地点相关的资源配置同样具有相关的联系，如院方的出诊情况、医疗设备、资源配置等等，如有这方面的数据，可以更进一步的探索能够够提高相关数据模型的特征方法来优化模型。